VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Assessing the Methodological Quality of Systematic Reviews

Shea, B.J.

2008

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Shea, B. J. (2008). *Assessing the Methodological Quality of Systematic Reviews: The Development of AMSTAR*. [PhD-Thesis – Research external, graduation internal, Vrije Universiteit Amsterdam].

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 05. May. 2023

Assessing the Methodological Quality of Systematic Reviews
The Development of AMSTAR

Beverley Julia Shea

ASSESSING THE METHODOLOGICAL QUALITY OF SYSTEMATIC REVIEWS

THE DEVELOPMENT OF AMSTAR

Beverley Julia Shea

VRIJE UNIVERSITEIT

Assessing the Methodological Quality of Systematic Reviews The Development of AMSTAR

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof. Dr. L.M. Bouter
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op 8 oktober 2008 om 13.45 uur
in de het auditorium van de universiteit,
De Boelelaan 1105

door

Beverley Julia Shea

geboren te St. John's, Newfoundland and Labrador, Canada

promotoren: prof.dr. L.M. Bouter
prof.dr. M. Boers
copromotor: prof.dr. J.M. Grimshaw

ISBN/EAN 978-90-5669-117-2
NUR-code 870

The study presented in this thesis was performed at the Institute for Research in Extramural Medicine (EMGO Institute) of the VU University Medical Center (VUmc), the Netherlands and the Department of Clinical Epidemiology and Biostatistics (KEB) of the VU University Medical Center, the Netherlands. The EMGO Institutes participates in the Netherlands School of Primary Care Research (CaRe), which was re-acknowledged in 2006 by the Royal Netherlands Academy of Arts and Sciences (KNAW).

CONTENTS

Chapter 1	General Introduction	6
Chapter 2	Assessing the quality of reports of systematic reviews: The QUOROM statement compared to other tools.	12
Chapter 3	Scope for improvement in the reporting quality of systematic reviews from the Cochrane musculoskeletal group.	29
Chapter 4	Does updating improve the methodological and reporting quality of review quality and the reporting quality of Cochrane reviews?	43
Chapter 5	Development of AMSTAR: a measurement tool to assess methodological quality of systematic reviews.	56
Chapter 6	Internal validation of AMSTAR: a measurement tool to assess systematic reviews.	66
Chapter 7	External validation of a measurement tool to assess systematic reviews (AMSTAR)	76
Chapter 8	Discussion	89
	Instruments	100
	Summary	108
	Samenvatting	111
	Acknowledgements	114
	About the author	119

1

1

GENERAL INTRODUCTION

A systematic review is a comprehensive assessment of the medical literature on a topic of interest using *a priori* specified rules for the search, identification and eligibility of the pertinent studies, and for the abstraction of relevant data.[1] The systematic nature of the process, which is carried out according to clear-cut rules, differentiates a systematic review from a traditional review authored by experts without the self-imposed discipline of specified rules. Due to the explosion of biomedical publishing in the latter half of the 20th century (perhaps 30,000 journals and upwards of two million primary research articles a year), keeping up with primary research is an impossible feat.[2] Systematic reviews have become essential tools of social and medical study. Patients and clinicians are making increasing use of systematic reviews in making evidence-based decisions about treatments.[3] Those who rely on systematic reviews assume that the methodologies of organizations such as the Cochrane Collaboration and the Campbell Collaboration are rigorously developed and therefore, the quality of their reviews is the best it can possibly be.[4,5] But how can the users of systematic reviews know whether their confidence is justified?

The Cochrane Collaboration, which is an international multi-disciplinary organization established in 1993, has the avowed task of preparing, maintaining, and disseminating systematic, up-to-date reviews of health care.[4] The purpose of conducting systematic reviews is to gain valid and reliable information that will guide evidence-based decisions. The issues reported are often complex, but when the evidence gathered is strong, and its implications clear, it is hoped that the review will influence decision making and help to shape health policy.

The production of comprehensive and accessible pre-appraised resources supports an evidence-based approach to decision making. Systematic reviews have added valuable information to the pool of resources.[6,7] The Cochrane Library now contains over 3,500 well-designed systematic reviews covering a variety of problems.[8] If systematic reviews are to be useful, serious consideration must be given to how they are conducted and reported.

The ultimate test of a systematic review is whether its report justifies confidence that it is evidence-based and that it accurately reflects the process followed during its various stages. One way to assess the merits of a systematic review is to examine the validity of its report. However, a systematic review may not reflect the manner in which its authors conducted their review so much as it does their ability to write comprehensibly.[9]

Prior to, and in the course of performing research for this thesis, several authors documented the considerable variations in the quality of published reviews. Silagy surveyed 28 systematic reviews published in primary care journals during 1991 using 8 methodological criteria understood to be important in the reporting of systematic reviews.[10] Each criterion had a maximum score of 2, for a total score of 16. Silagy reported that only 25% of these systematic reviews obtained a total score higher than 8. Recently, Moja et al. assessed 965 systematic reviews published between 1995 and 2002 in the Cochrane Library and in paper-based journal formats. They concluded that the reviews failed to take several important factors into account in their interpretation of results; that methods for assessment of methodological quality by systematic review are still in their infancy; and that there is substantial room for improvement.[11]

Silagy et al. studied the distinction between methodological quality and reporting quality. He found evidence that researchers deviated from their original study protocol during the execution of the review process without clearly documenting same deviations in the final study report.[12] Liberati et al. found little distinction between methodological quality and reporting quality.[13] Findings such as those reported above have lead to the conduct of studies to assess methodological quality and the development of instruments such as the quality of reporting of meta-analysis statement (QUOROM). Such initiatives should encourage improvement in the reporting quality of systematic reviews.[14,15]

As we experienced in the course of the studies described in this thesis, there are important differences between assessing the methodological quality of systematic reviews and assessing their reporting quality.[16-18] The first, *methodological quality*, considers how well the systematic review was conducted (literature searching, pooling of data, etc.). The second, *reporting quality*, considers how well systematic reviewers have reported their methodology and findings.

In summary, systematic reviews have become an integral part of scholarly practice. However, the method to assess their quality is not fully developed.

This thesis was comprised of two sets of three related research projects, each guided by its own objective and reported in its own Chapter.

Objective 1: *To review the current status of instruments used to assess the reporting quality of systematic reviews (Chapter 2)*

To select the most appropriate instrument for further use, we conducted a study to compile and appraise a complete list of all available tools for the assessment of systematic reviews. We improved the descriptors of the instrument that came out on top (*overview quality assessment questionnaire, OQAQ*).

Objective 2: *To assess the reporting quality of a complete subset of electronic systematic reviews, by applying OQAQ and QUOROM (Chapter 3)*

We applied the enhanced OQAQ and QUOROM to all 57 Cochrane Musculoskeletal (CMSG) systematic reviews published in the Cochrane Database of Systematic Reviews of the Cochrane Library, Issue 4, 2002.

Objective 3: *To determine the impact of updating on the methodological quality and reporting quality of a subset of systematic reviews (Chapter 4)*

Under this objective we assessed a newly selected sample of updated systematic reviews before and after their updating using the same two instruments. The sample covered a wide variety of health topics published in the Cochrane Library. This exercise provided a second test of the applicability of the instruments used under objective 2.

We concluded there was room for a new instrument focused on methodological quality (rather than reporting quality) of systematic reviews, with improved content and feasibility. The next 3 Chapters describe the development and validation of this new instrument, termed AMSTAR, an acronym for "**A M**ea**S**urement **T**ool to **A**ssess Systematic **R**eviews".

Objective 4: *To develop a valid and reliable quality assessment instrument for systematic reviews (Chapter 5)*

AMSTAR was developed from a comprehensive set of possible items retrieved from the OQAQ and a

comprehensive list drawn up by Sacks, to which three new items/dimensions were added. All items were scored in a large dataset of reviews, and these scores were subjected to factor analysis. An international panel of experts appraised the resulting domains and selected the best item per domain in a nominal group consensus process.

Objective 5: *To test the validity and reliability of AMSTAR in the source dataset (Chapter 6)*

We tested the new instrument by having two assessors apply it and the two original instruments to a random sample of 30 systematic reviews (out of the 151 selected under objective 4). The purpose of this exercise was to compare the validity and reliability (and feasibility) of the new instrument to that of the two existing instruments.

Objective 6: *To externally test the validity and reliability of AMSTAR (Chapter 7)*

We conducted a second validation study to test the reliability of AMSTAR using a separate set of reviews. External assessors naive to AMSTAR applied the instrument to a set of 42 reviews assessing the use of protein pump inhibitors for gastroesophageal reflux disease, dyspepsia and peptic ulcer disease. The following table summarizes the various measurement instruments used in the course of our research, the sets of systematic reviews we used in this study, and the systematic reviews to which we applied each instrument (Table 1).

Table 1: Measurement instruments and reviews studied

	OQAG[1]	QUOROM[2]	Sacks[3]	AMSTAR[4]	Global Assessment	Set of reviews
Chapter 2: General Systematic Reviews	X	X	X			4 systematic reviews
Chapter 3: CMSG[5]	X	X				57 CSMG reviews
Chapter 4: Updated Reviews	X	X				53 Cochrane reviews (assess-ments-pre and update)
Chapter 5: AMSTAR Development	X		X +3 items			Factor analysis derived from a dataset of 151 systematic reviews
Chapter 6: AMSTAR Validation 1	X		X	X		30 systematic reviews ran-domly selected from the above sample of 151 systematic reviews
Chapter 7: AMSTAR Validation 2				X	X	42 systematic reviews on gastroenterology interventions

1. Enhanced overview quality assessment questionnaire (OQAQ)
2. Quality of reporting of meta-analysis (QUOROM)
3. Sacks (developed by Henry Sacks et al.)
4. A measurement tool to assess systematic reviews (AMSTAR)
5. Cochrane musculoskeletal group (CMSG)

Research done on each of the objectives is described in an article that has been published in, or submitted to, a scientific journal. Consequently, each Chapter can be read independently. Because of the inter-related nature of the objectives, some degree of overlap in introduction and methods sections could not be avoided. In some instances, an addendum has been added to accommodate insights gained after publication of these Chapters.

A general discussion of our overall project is provided following the seven main Chapters of this thesis. The discussion describes the major findings for each of the objectives listed in this introduction, outlines some of the challenges faced in our research, and provides recommendations for future application and further research **(Chapter 8)**.

We invite you to follow us as we describe the journey that lead to the development of AMSTAR.

References

1. Chalmers I, Altman DG. Systematic Reviews. London: BMJ Publications 1995.
2. Davies HT, Crombie IK. What is a systematic review? 2003: Hayward Medical Communications www.evidence-based-medicine.co.uk.
3. Tugwell P, Shea B, Boers M, Brooks P, Simon L, Strand V, Wells G. Evidence Based Rheumatology. *BMJ Books* 2004.
4. Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating reviews of the effects of health care. JAMA 1995 Dec; 274(24): 1935-38.
5. The Campbell Collaboration http://www.campbellcollaboration.org/.
6. Cook DJ, Mulrow CD, Haynes RB. Systematic Review: Synthesis of Best Evidence for Clinical Decisions. *Ann Intern Med* 1997 Mar; 126(5): 376-80.
7. Mulrow CD, Cook DJ, Davidoff F. Systematic Review: Critical Links in the Great Chain of Evidence. *Ann Intern Med* 1997 Mar; 126(5): 389.
8. *The Cochrane Library*, Issue 3, 2007. Chichester, UK: John Wiley & Sons, Ltd.
9. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of randomized controlled trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999; 3(12): i-iv, 1-98.
10. Silagy C. An analysis of review articles published in primary care journals. *Fam Pract* 1993 Sept; 10(3): 337-41.
11. Moja L, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A and on behalf of the Metaquality Study Group. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. *BMJ* 2005 May; 330(7499): 1053.
12. Silagy C, Middleton P, Hopewell S. Publishing protocols of systematic reviews: Comparing what was done to what was planned. *JAMA* 2002 Jun; 287(21): 2831-34.
13. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol* 1986 Jun; 4(6): 942-51.
14. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. *Lancet* 1999 Nov; 354(9193): 1896-1900.
15. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. Systematic Review in Health Care Meta-analysis in context. *BMJ Books* 2001: 122-39.
16. The AGREE Collaboration. Writing Group: Cluzeau FA, Burgers JS, Brouwers M, Grol R, Mäkelä M, Littlejohns P, Grimshaw J, Hunt C. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Quality and Safety in Health Care* 2003; 12(1): 18-23.
17. Smidt N, Rutjes A, van der Windt D, Ostelo R, Reitsma J, Bossuyt P, Bouter LM, de Vet H. Quality of Reporting of Diagnostic Accuracy Studies. *Radiology* 2005 May; 235(2): 347-53.
18. Moher D, Soeken K, Sampson M, Campbell K, Ben Perot L, Berman B. Assessing the quality of reports of systematic reviews in pediatric complementary and alternative medicine. *BMC Pediatr* 2002; 2:3.

2

2

ASSESSING THE QUALITY OF REPORTS OF SYSTEMATIC REVIEWS: THE QUOROM STATEMENT COMPARED TO OTHER TOOLS

Systematic reviews within health care are conducted retrospectively which makes them susceptible to potential sources of bias. In the last few years, steps have been taken to develop evidence based methods to help improve the reporting quality of randomized trials in the hope of reducing bias when trials are included in meta-analysis. Similar efforts are now underway for reports of systematic reviews.

This paper describes the development of the QUOROM statement and compares it to other instruments identified through a systematic review. There are many checklists and scales available to be used as evaluation tools, but most are missing important evidence based items when compared against the QUOROM checklist.

A pilot study suggests considerable room for improvement in the quality of reports of systematic reviews using four different instruments. It is hoped that journals will support the QUOROM statement in a similar manner to the CONSORT statement.

Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: The QUOROM statement compared to other tools. In: Egger M, Smith GD, Altman DG, eds. Systematic Reviews in Health Care. Meta-analysis in context. *BMJ Books* 2001:122-139.

Introduction

There are approximately 17,000 biomedical books published every year and 30,000 biomedical journals, with an annual increase of 7%.[1] This makes it very difficult for health care professionals to stay apprised of the most recent advances and research in their respective fields as they would be required to read an average of 17 original articles each day.[2]

To make this task slightly more manageable, health care providers and other decision-makers now have among their information resources access to a form of clinical report called the systematic review. This is a review in which bias has been reduced by the systematic identification, appraisal, synthesis, and if relevant, statistical aggregation of all relevant studies on a specific topic according to a predetermined and explicit methodology. Theoretically, such reviews can effectively summarize the accumulated research on a topic, promote new questions on the matter, and channel the stream of clinical research towards relevant horizons. Consequently, systematic reviews can also be important to health policy planners and others involved in planning effective health care.

If the results of systematic reviews are to be used by health care providers and health care consumers, it is necessary that they are as uninhibited of bias as possible (i.e., systematic error). One way to assess the merits of a systematic review is to assess the quality of its report. It is possible that a scientific report may not reflect how the investigators conducted their review but rather, their ability to write comprehensively. Although the data addressing this point is sparse, it appears that a scientific report is a reasonable marker as to how the project was conducted. In an assessment of the quality of 63 randomized trials in breast cancer, Liberati and colleagues[3] reported that the average quality of reports was 50% (95%CI: 46 to 54 %). Following these assessments, the investigators interviewed 62 of the corresponding authors to ascertain whether information in the manuscripts submitted to publication consideration was removed prior to its publication. The authors reported that with the additional information obtained from the interviews, the quality scores only increased marginally to an average score of 57%. These data come from clinical trials. We are unaware of comparable data for systematic reviews.

Choosing an appropriate evaluation tool for critically appraising the report of a systematic review is as difficult as the assessment of the quality of reports of randomized trials. A systematic review[4] designed to identify and appraise instruments that assess the quality of reports of randomized trials found twenty-five scales and nine checklists. The scales differed considerably from one another in a variety of areas including: how they defined quality; the scientific rigor in which they were developed; the number of items they used; and the time required to use them. When six of the scales were compared to one another to assess the same randomized trials, divergent scores and rankings were reported.

In an attempt to attain consistency in the reporting quality, the purpose of this Chapter is to identify and appraise instruments developed to assess the quality of reports of systematic reviews. It will also evaluate whether different instruments assessing the same meta-analysis would provide similar evidence regarding its quality.

A systematic review of published checklists and scales

A literature search was performed to take an inventory of published checklists and scales. Potentially relevant articles were chosen and the tools described were reviewed. Quality assessment, across a sample of conveniently selected instruments was tested based on four randomly chosen systematic reviews. A more detailed description of this process can be found in Box 2: Methodology.

METHODOLOGY

Literature Search

- MEDLINE: January 1966- February 1999
- three independent searches with keywords: meta-analysis, review literature, systematic or quantitative or methodologic review, overview, review, information synthesis, integrative research review, guideline, checklist, tool, scoring, scale, clinimetric, quality, critical reading, methodology.
- PubMed "related articles" function to find others

Identification and selection

- initial screening to identify relevance
- potentially relevant articles reviewed independently by each author
- article eligible regardless of language
- article has to be scale or checklist (def. Annex I) designed to assess quality of systemic reviews and meta-analyses

Data Extraction

- checklists and scales assessed for: 1) number of items included in tool, 2) aspects of quality assessed, 3) whether or not article included explicit statement regarding purpose of tool, and 4) time for completion of tool
- data extraction was completed in a group and a consensus was reached

Quality assessment

- compared items in each quality assessment instrument against QUOROM statement
- three checklists and one scale were conveniently selected to compare stability of quality assessments across instruments
- randomly selected four systematic reviews (from pool of 400 systematic reviews) to be used for quality assessment based on four selected instruments
- quality assessments completed as a group
- quality assessment established in two ways: 1) a quantitative estimate based on one item from a validated scale and 2) the proportion of items reported (as a function of the number of items included in tool) in the systematic review report

The QUOROM Statement

Although guidelines for reporting systematic reviews have been suggested, a consensus across disciplines as to how they should be reported has only recently been developed. Following a recent initiative to improve the reporting quality of randomized controlled trials[5], a conference referred to as the Reporting quality of Meta-analyses (QUOROM) was held to address these issues as they relate to systematic reviews of randomized trials. The QUOROM conference participants included clinical epidemiologists, clinicians, statisticians, researchers who conduct meta-analysis, and editors from the United Kingdom and North America who were interested in systematic reviews. This conference resulted in the creation of the QUOROM Statement, which consists of a checklist (Instrument 1) and flow diagram (Instrument 1).[6]

The checklist consists of 18 items, including 8 that are evidence-based[7-20], which address primarily the Abstract, Introduction, Methods, and Results section of a report of a systematic review of randomized trials. This checklist encourages authors to provide readers with information regarding searches, selection, validity assessment, data abstraction, study characteristics, quantitative data synthesis, and trial flow. The flow diagram provides information about the progress of randomized trials throughout the review process from the number of potentially relevant trials identified to those retrieved and ultimately included. Items reported in the QUOROM statement that are to be included in a systematic review report were chosen based on evidence whenever possible, which implies the need to include items that can systematically influence estimates of treatment effects.

Over the last eighteen months, the QUOROM group has evaluated the impact of the QUOROM statement on the editorial process. Ten medical journals have participated in a randomized trial to evaluate the impact of applying the QUOROM criteria on journal peer review. Accrual is now complete and the trial results should be reported soon.

QUOROM

The QUOROM statement for reporting systematic reviews was created according to evidence and provides a comprehensive set of guidelines. Several methods were used to generate the checklist and flow diagram. These included a systematic review of the reporting of systematic reviews; focus groups of the Steering Committee; and a modified Delphi approach during an expert panel consensus conference. QUOROM group members were asked to identify items that they thought should be included in a checklist that would be useful for investigators, editors, and peer reviewers. The inclusion of items in the checklist was guided by research evidence suggesting that a failure to adhere to the particular checklist item proposed could lead to biased results.

For example, authors are asked (under the 'Methods' heading and 'Searching' subheading) to be explicit in their reporting as to whether or not they have used any restrictions on language of publication. Approximately one-third of published systematic reviews have some language restrictions as part of the eligibility criteria for including individual trials.[11] It is not clear why this is done since there is no evidence to support differences in study quality[12] and there is evidence supporting the view that such action may result in a biased summary. The role of language restrictions has been studied in 211 randomized trials included in 18 meta-analyses in which trials published in languages other than English were included in the quantitative summary.[11] Language-restricted meta-analyses, as compared to language-inclusive ones, overestimated the treatment effect by only 2%, on average. However, the language-inclusive meta-analyses were more precise.[11]

The QUOROM statement was also formally pre-tested with representatives of several constituencies who would use the recommendations. Modifications were made based on pre-test results.

Results

The search identified 318 potentially relevant articles. After eliminating duplicates or previously published instruments and those that were not scales or checklists, 26 instruments[21-44] were included in our review. These included 23 checklists and 3 scales (**Table 1**, see pages 22 and 23).

Table 1
Number of criteria reported by each checklist and scale (first author named) fulfilling the 18 headings and subheadings included in the QUOROM statement

Instrument	ABSTRACT						INTRODUCTION	METHOD						RESULTS			DISCUSSION
		Objectives	Data sources	Review methods	Results	Conclusion		Searching	Selection	Validity assessment	Data abstraction	Desc. of study Characteristics	Quantitative Data synthesis	Trial flow	Desc. of study Characteristics	Quantitative Data synthesis	
Checklist																	
BLETTNER[21]	N	N	N	N	N	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y	N
COOK[22]	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y
GELLER[23]	N	N	N	N	N	N	N	Y	Y	Y	N	N	Y	N	N	Y	Y
GOLDSCHMIDT[24]	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	N	Y	Y
GREENHALGH[25]	N	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	N	Y
HALVORSEN[26]	N	N	N	N	N	N	N	Y	Y	N	N	Y	Y	N	N	N	N
IRWIG[27]	N	N	N	N	N	N	Y	Y	Y	Y	Y	U	Y	N	Y	Y	Y
L'ABBÉ[28]	N	N	N	N	N	N	U	Y	Y	Y	Y	Y	Y	N	N	U	Y
LIGHT[29]	N	N	N	N	N	N	Y	N	Y	N	N	Y	N	N	N	N	U
MEINERT[30]	Y	Y	Y	Y	N	Y	N	Y	Y	N	N	Y	Y	N	Y	Y	Y
MULLEN[31]	N	N	N	N	N	N	U	Y	Y	Y	Y	N	Y	N	U	Y	Y

MULROW[32]	N	N	N	N	N	N	Y	Y	Y	Y	N	N	N	N	Y	Y	U
NEELY[33]	N	Y	Y	U	U	Y	Y	Y	Y	Y	N	U	U	N	U	U	Y
NONY[34]	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
OHLSSON DIAG[35]	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	U
OHLSSON RCT[36]	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	U
OXMAN 1994[36]	N	N	N	N	N	N	N	Y	Y	Y	N	N	N	N	N	N	Y
OXMAN CMAJ[37]	N	N	N	N	N	N	N	Y	Y	Y	Y	U	U	N	Y	Y	N
POGUE[38]	N	N	N	N	N	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y	N
SACKS[39]	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
SMITH[40]	N	N	N	N	N	N	Y	N	Y	Y	N	Y	N	N	U	N	U
THACKER[41]	N	N	N	N	N	N	N	Y	Y	Y	Y	U	U	N	U	U	Y
WILSON[42]	N	N	N	N	N	N	Y	Y	Y	Y	Y	U	Y	N	Y	Y	N
Scale																	
ASSENDELFT[43]	N	N	N	N	N	N	N	Y	Y	Y	N	Y	N	N	Y	Y	Y
AUPERIN[44]	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	N	U	Y	Y
Enhanced OQAQ[21] (instrument[2])	N	N	N	N	N	N	N	Y	Y	Y	Y	N	Y	N	N	Y	Y

Each of these instruments has now been published. The instruments were developed between 1984 and 1997, indicating a fairly recent interest in this area. The number of items in each instrument ranged from 5 to 101, with only two checklists having more than 35 items (**Table 2**).[22, 24] Three checklists and one scale were developed to assess quality assessment within specific domains, such as diagnostic tests (**Table 2**). The remaining 20 checklists and two scales were developed for use with all types of systematic reviews. Fifteen checklists and one scale included an explicit statement regarding the purpose of the instrument. The average time required to assess the quality of a systematic review using the checklists and scales was 13 minutes (range: 5 to 30) and 12 minutes (range: 5 to 20) respectively (**Table 2**). Two of the checklists took at least 30 minutes to complete.[22, 24]

Table 2
Descriptive characteristics of published checklists and scales used to assess the quality of systematic reviews and meta-analyses of randomized trials

Instrument	Number of items	Type of quality assessed	Explicit statement regarding the purpose of tool	Time to complete*
Checklist				
BLETTNER[21]	Y	Y	N	Y
COOK[22]	Y	Y	Y	Y
GELLER[23]	N	N	Y	N
GOLDSCHMIDT[24]	Y	Y	Y	Y
GREENHALGH[25]	N	N	Y	N
HALVORSEN[26]	N	N	N	N
IRWIG[27]	Y	Y	Y	Y
L'ABBÉ[28]	U	U	Y	U
LIGHT[29]	Y	Y	U	Y
MEINERT[30]	N	N	Y	N
MULLEN[31]	U	U	Y	U
MULROW[32]	Y	Y	U	Y
NEELY[33]	Y	Y	Y	Y
NONY[34]	Y	Y	Y	Y
OHLSSON DIAG[35]	Y	Y	U	Y
OHLSSON RCT[36]	Y	Y	U	Y
OXMAN 1994[36]	N	N	Y	N
OXMAN CMAJ[37]	N	N	N	N
POGUE[38]	Y	Y	N	Y
SACKS[39]	N	N	Y	N
SMITH[40]	Y	Y	U	Y
THACKER[41]	N	N	Y	N
WILSON[42]	Y	Y	N	Y
Scales				
ASSENDELFT[43]	N	N	Y	N
AUPERIN[44]	N	N	Y	N
Enhanced OQAQ[21] (instrument[2])	N	N	Y	N

* Approximate time which may vary depending on the operator
** There are several sub categories within each of the questions

Comparison of QUOROM to other checklists and scales

Checklists

None of the other checklists included all the items recommended by QUOROM (see **Table 1**). The majority of checklists contained items about what the methods section of a systematic review should include and generally neglected the other components of the report: only one (4%) checklist[22] included an item regarding the title and two (11%) addressed the abstract.[30, 33] The Abstract items in the QUOROM checklist were the least frequently encountered among the checklists (0% to 9%). Thirteen (56%) included an item about the introduction, although we could not determine whether this criterion was met in two[28, 31] of the checklists.

There was considerable overlap between the content of the QUOROM checklist and the methods section of the other checklists. All but two checklists (91%) questioned searching methods and all but one (96%) questioned selection criteria. Eighteen (78%) included an item on validity and fourteen (61%) asked about data abstraction. Items concerning data synthesis were definitely present in 13 checklists (68%) and were possibly present in three others (see **Table 1**). However, while quantitative data synthesis was clearly identified as a prerequisite in 15 (65%) of the checklists and may possibly have been required in three others, only nine (39%) of them (and possibly five others) contained a question on the characteristics of individual studies in the methods section.

Items concerning the results and discussion sections were definitely reported in 69% and 61% of the cases respectively, with the exception of the "trial flow" item which was not included in any of the checklists. Thirteen checklists (56%) and possibly four others stressed the need for its inclusion in the results section. Again, the need for quantitative data synthesis in the results section was mentioned in the majority of checklists, i.e. 15 (65%), and possibly in three others. Thirteen checklists (56%) and possibly five others included an item about a discussion section (see **Table 1**).

Scales

Unlike the QUOROM statement, none of the scales included a question on the title, the introduction, or the abstract. The Abstract items in the QUOROM checklist were the least frequently encountered among the scales (0%), while those concerning the methods sections were the most frequently encountered, i.e. from 42% to 96% of the time. For example, all three scales included items on searching, selection, and validity assessment. Data abstraction, describing the study characteristics of the primary studies and quantitative data synthesis were included in two of the three scales (see **Table 1**). In the results section, no scale suggested producing a trial flow diagram as recommended by the QUOROM statement. All scales included a quantitative data synthesis and three of them[36, 44, Instrument 1] included an item on describing the characteristics of the primary studies. The scales also included an item about the discussion.

Assessment of quality across instruments

In order to compare the stability of quality assessments across instruments, we completed a small pilot study. We conveniently selected three checklists from Table 1[28, 30, 39] and one scale[Instrument 1], representing a broad spectrum. Out of the four systematic reviews evaluated using the checklists and scales, three were paper based[45-47] while the fourth was a Cochrane review [48]. All four were published in 1992 or later (**Table 3**). The quality of the report of each review, based on the proportion of items reported in the review, was fairly stable across instruments. The difference in quality scores, between the four instruments ranged from 26% to 34% of the maximum possible evaluation (**Table 3**). The Cochrane review had the highest

quality regardless of the instrument used. When the systematic reviews were ranked based on the quality score obtained from each instrument, it was apparent that one of them consistently ranked highest, independent of the instrument. The rank ranges were also stable across the different instruments used, with the Cochrane review reporting the highest rank across all five instruments (**Table 3**).

Table 3
Quality scores and ranks of each of four meta-analyses of randomized trials, across four quality assessment instruments

Title of Meta-analysis	Year of Publication	Original Oxman & Guyatt/OQAQ	Enhanced OQAQ Oxman & Guyatt	L'Abbé	Sacks	Meinert	% Range (Difference)	Rank range
		Score	Proportion of Items present % (number of items/total); rank					
Effects of psychosocial interventions with adult cancer patients: a meta-analysis of randomized experiments[46]	1995	43 (3/7); 2	56 (5/9); 2	22 (2/9); 2	22 (2/9); 2	49 (17/35); 2	22-56 (34)	2-3
Impact of post menopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials[47]	1997	29 (2/7); 3	22 (2/9); 4	11 (1/9); 4	26 (6/33); 3	37 (13/35); 4	11-37 (26)	3-4
How effective are current drugs for Crohn's Disease? A meta-analysis[48]	1992	29 (2/7); 3	56 (5/9); 2	22 (2/9); 2	35 (8/33); 2	43 (15/35); 3	22-56 (34)	2-3
The efficacy of tacrine in Alzheimer's Disease[49]	1997	57 (4/7); 1	67 (6/9); 1	33 (3/9); 1	65 (15/23); 1	49 (17/35); 1	33-67 (34)	1

Discussion

The literature search yielded more than two dozen instruments that had been developed to assess various aspects of a systematic review. In our view, the enhanced scale [Instrument 2] met several important criteria. Its original developers defined the construct they were interested in investigating, measured the discriminatory power of items, and conducted inter observer reliability studies as part of the development process.[36]

A full discussion of the process of instrument development is beyond the scope of this Chapter. However, it is fair to say that investigators wishing to develop a new instrument should, at the very least, follow published guidelines for instrument development.[49, 50]

Regardless of the checklist or scale used, the results of this assessment indicate that the quality of reports of meta-analysis is low and there is considerable room for improvement. Similar data has been reported elsewhere. A classic 1987 survey[39] of 86 meta-analyses assessed each publication on 14 items from six content areas thought to be important in the conduct and reporting of meta-analysis of randomized trials: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. The results showed that only 24 (28%) of the 86 meta-analyses addressed all six content areas. When updated, using more recently published systematic reviews, our survey yielded similar results.[51] Comparable results have also been reported elsewhere.[52, 53]

These results indicate that not only systematic reviewers, but also editors and peer reviewers, may not fully appreciate the elements that should be taken into account during the design, conduct, and reporting of reviews.

The exception to the low quality of reports was the quality of the Cochrane review.[48] It was found that the Cochrane review had the highest absolute score and rank regardless of the instrument used. These results provide further evidence suggesting that the quality of reports of Cochrane reviews are of higher quality, in some aspects, than paper-based ones. In a review of 36 Cochrane reviews, compared to 39 paper-based ones published in 1998, the authors found that Cochrane reviews, unlike paper-based ones, included a description of the inclusion and exclusion criteria and assessed trial quality.[54] In addition, Cochrane reviews were more frequently updated and none had language restrictions as eligibility criteria.

The majority of instruments designed to assess the quality of systematic reviews are checklists rather than scales. Among the checklists, there is considerable range in the number of items and time required for completion. The scales are somewhat more consistent in terms of the item pool and time required for use. The instruments have been developed to assess systematic reviews in several content areas: psychology, internal medicine, surgery, and rehabilitation health. Most instruments have been developed to assess systematic reviews of randomized trials, although we found two checklists specifically designed for diagnostic tests.[27, 35]

Results reported in this Chapter suggest that the items included in the checklists and scales focus particularly on aspects of the methodology of systematic reviews. Items in the QUOROM checklist were also heavily weighted in the methods section. However, items in the QUOROM also gave close scrutiny to the abstract while many of the other instruments had next to no items addressing the abstract. Given that many readers only read the abstract, it was disappointing to see so little attention paid to this aspect of reports.

With exception to QUOROM, no instrument was found that asked authors to report on the flow of studies through the various stages of a systematic review. Such information regarding the process used by the authors to include studies throughout the review process has clear face validity for readers. Some items included in other assessment tools are not covered within the QUOROM checklist; but given the methodology used to develop the QUOROM we do not believe that many evidence-based criteria were missed.

If the quality of reports of systematic reviews is to improve, steps must be taken. The Cochrane Collaboration is addressing this objective through a combination of continual peer review throughout the

systematic review process and the use of strict criteria that must be included in the process and report. The use of evidence-based criteria, such as those identified in the QUOROM statement, should also help to improve the situation. Similar efforts have recently been made to help improve the quality of reports of randomized trials and we are starting to see the benefits of this approach.[55] We hope that journals that have endorsed the CONSORT statement will do the same for the QUOROM statement.

References

1. Smith R. Where is the wisdom...?: the poverty of medical evidence. *BMJ* 1991 Oct; 303(6806): 798-99.
2. Davidoff F, Haynes B, Sackett D, Smith R. Evidence-based medicine: a new journal to help doctors identify the information they need. BMJ 1995 Apr; 310(6987): 1085-86.
3. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol* 1986 Jun; 4(6): 942-51.
4. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technol Assess Health Care* 1996 Spring; 12(2): 195-208.
5. Begg CB, Cho MK, Eastwood S, Horton R, Moher D, Olkin I, Rennie D, Schulz KF, Simel DL, Stroup DF. Improving the reporting quality of randomized controlled trials: the CONSORT statement. *JAMA* 1996 Aug; 276(8): 637-39.
6. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup D for the QUOROM group. Improving the reporting quality of meta-analysis of randomized controlled trials: the QUOROM statement. *Lancet* 1999 Nov; 354(9193): 1896-900.
7. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994 Nov; 309(6964): 1286-91.
8. Taddio A, Pain T, Fassos FF, Boon H, Illersich AL, Einarson TR. Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. *CMAJ* 1994 May; 150(10): 1611-15.
9. Tramér M, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997 Sept; 315(7109): 635-40.
10. McAuley L, Moher D, Tugwell P. The influence of 'grey' literature on meta-analysis. Technical report. Thomas C. Chalmers Centre for Systematic Reviews. 1999.
11. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, Liberati A. Does the language of publication of reports of randomized trials influence the estimates of intervention effectiveness reported in meta-analyses? Cochrane Colloquium 1998.
12. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997 Aug; 350(9074): 326-29.
13. Khan KS, Daya S, Collins JA, Walter S. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* 1996 May; 65(6): 939-45.
14. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995 Feb; 273(5): 408-12.
15. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does the quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998 Aug; 352(9128): 609-13.
16. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996 Feb; 17(1): 1-12.
17. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-Analysis Blinding Study Group. *Lancet* 1997 Jul; 350(9072): 185-86.
18. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* 1998 May; 279(19): 1566-70.
19. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994 Nov; 309(6965): 1351-55.

20. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986 Oct; 4(10): 1529-41.
21. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol* 1999 Feb; 28(1): 1-9.
22. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis. *J Clin Epidemiol* 1995 Jan; 48(1): 167-71.
23. Geller NL, Proschan M. Meta-analysis of clinical trials: a consumer's guide. *J Biopharm* Stat 1996 Nov; 6(4): 377-94.
24. Goldschmidt PG. Information Synthesis: A practical guide. *Health Serv Res* 1986 Jun; 21(2 Pt 1): 215-37.
25. Greenhalgh T. Papers that summarise other papers (systematic reviews and meta-analyses). In: How to read a paper - the basics of evidence based medicine. BMJ Publishing Group, London. 1997.
26. Taylor Halvorsen K. The reporting format. Edited by Cooper H and Hedges LV. 1994: 425-37.
27. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995 Jan; 48(1): 119-30.
28. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987 Aug; 107(2): 224-33.
29. Light RJ, Pillemer DB. The Science of reviewing research. *Harvard Univ Press* 1984; 160-86.
30. Meinert CL. Meta-analysis: Science or religion? *Control Clin Trials* 1989 Dec; 10(4 Suppl): 257S-263S.
31. Mullen PD, Ramirez G. Information synthesis and meta-analysis. *Advances in Health Education and Promotion* 1987; 2: 201-39.
32. Mulrow CD. The medical review article: state of the science. *Ann Intern Med* 1987 Mar; 106(3): 485-88.
33. Neely JG. Literature review articles as a research form. *Otolaryngol Head Neck Surg* 1993 Jun; 108(6): 743-48.
34. Nony P, Cucherat M, Haugh MC, Boissel JP. Critical reading of the meta-analysis of clinical trials. *Therapie* 1995 Jul-Aug; 50(4): 339-51.
35. Ohlsson A. Systematic reviews - theory and practice. *Scand J Clin Lab Invest* 1994; 219: 25-32.
36. Oxman AD. Checklists for reviews articles. *BMJ* 1994 Sep; 309(6955): 648-51.
37. Oxman AD, Guyatt G. Guidelines for reading literature reviews. *CMAJ* 1988 Apr; 138(8): 697-703.
38. Pogue JM and Yusuf S. Overcoming the limitations of current meta-analysis of randomized controlled trials. *Lancet* 1998 Jan; 351(9095): 47-52.
39. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *NEJM* 1987 Feb; 316(8): 450-54.
40. Smith MC, Stullenbarger E. Meta-analysis: an overview. *Nurs Sci* Q 1989 Fall; 2(3): 114-15.
41. Thacker SB, Peterson HB, Stroup DF. Meta-analysis for the obstetrician-gynecologist. *Am J Obstet Gynecol* 1996 May; 174(5): 1403-07.
42. Wilson A, Henry DA. Meta-analysis. Part 2: Assessing the quality of published meta-analyses. *Med J Aust* 1992 Feb; 156(3): 173-87.
43. Assendelft WJJ, Koes B, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA* 1995 Dec; 274(24): 1942-48.
44. Auperin A, Pignon JP, Poynard T. Critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. *Almeth Pharmacol Therapy* 1999; 11: 215-25.
45. Meyer TJ, Mark MM. Effects of psychosocial interventions with adult cancer patients: A meta-analysis of randomized experiments. *Health Psychol* 1995 Mar; 14(2): 101-08.
46. Hemminki E, McPherson K. Impact of postmenopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials. *BMJ* 1997 Jul; 315(7101): 149-53.

47. Salomon P, Kornbluth A, Aisenberg J, Janowitz HD. How effective are current drugs for Crohn's disease? A meta-analysis. *J Clin Gastroenterol* 1992 Apr; 14(3): 211-15.
48. Qizilbash N, Birks J, Arrieta JL, Lewington S, Szeto S. Tacrine in Alzheimer's disease. *Cochrane Database of Systematic Reviews* 2000(2).
49. Norman DL, Streiner D. Health Measurement scales - A practical guide to their development and use. 2nd Ed. Oxford University Press. Oxford 1995.
50. McDowell I, Newell C. Measuring Health: a guide to rating scales and questionnaires, 2nd edn. Oxford: Oxford University Press 1996.
51. Sacks HS, Reitman D, Pagano D, and Kupelnick B. Meta-analysis: an update. *Mt Sinai J Med* 1996 May-Sep; 63(3-4): 216-24.
52. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. *J Clin Epidemiol* 1996 Feb; 49(2): 235-43.
53. McAlister FA, Clark HD, van Walraven C, Straus SE, Lawson FME, Moher D, Mulrow C. The medical review article revisited: has the science improved? *Ann Intern Med* 1999 Dec; 131(12): 947-51.
54. Jadad AR, Cook DJ, Jones A, Klassen T, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: A comparison of cohcrane reviews with articles published in paper-based journals. *JAMA* 1998 Jul; 280(3): 278-80.
55. Lebeau DL, Steinmann WC, Patel KV. Has the randomized controlled trial literature improved after CONSORT? Presented to the 50 years of randomized trials BMJ 1998.

Annex 1

Checklists: provide a qualitative estimate of the overall quality of a systematic review using itemized criteria for comparing the reviews. Consequently, checklist items do not have numerical scores attached to each item.

Scales: are similar to checklists except that each item of a scale is scored numerically and an overall quality score is generated. To be considered a scale an instrument should measure over a continuum and provide an overall summary score.

ADDENDUM

CHAPTER 2

We used the following methodological framework as the basis for this work. 'When the quality of a systematic review is examined, two major aspects are assessed. The first, methodological quality, considers how well the systematic review was conducted (literature searching, pooling of data, etc.). The second, reporting quality, considers how well systematic reviewers have reported their methodology and findings.

The instruments chosen were selected conveniently on the basis of a specific set of criteria. Our primary objective was to test a sample of different types of measurement instruments on a sample of reviews. The authors selected a checklist, a scale, and the newly developed quality of reporting and enhanced methodological instruments. We selected a random sample of studies. We should have selected the instruments randomly.

We found that the overview quality assessment questionnaire (OQAQ) had been rigorously developed, yet we decided to improve the descriptors of the items therein which resulted in the 'enhanced OQAQ' that we have applied in the next Chapters.

In addition, but not explicitly mentioned in the published version of the Chapters, we chose the Sacks' instrument out of the list of possible instruments because it was very comprehensive, it came out high in the quality ranking (Chapter 2), and it had been used in a classic survey of methodological quality of systematic reviews over time.

3

3

SCOPE FOR IMPROVEMENT IN THE REPORTING QUALITY OF SYSTEMATIC REVIEWS FROM THE COCHRANE MUSCULOSKELETAL GROUP

The objective was to assess the reporting quality in Cochrane musculoskeletal systematic reviews (excluding back and injury reviews).

This study assessed all the Cochrane musculoskeletal group's systematic reviews in Issue 4, 2002, of the Cochrane Library Database of Systematic Reviews. Two reviewers independently extracted data and assessed quality. Two assessment tools were used, including an 18 item checklist and flow chart developed by the reporting quality of meta-analysis (QUOROM) consensus group, a 10 item scale, and the enhanced Oxman-Guyatt overview quality assessment questionnaire (OQAQ). One question on the latter scale (item 10) scores overall quality on a 7-point scale, with high scores indicating superior quality. We analyzed data using univariate approaches.

The 57 systematic reviews assessed were found to have good overall quality, with scores on individual items revealing only minor flaws. Documenting the flow of included and excluded studies and summarizing the results are two areas needing improvement in reporting. According to the Oxman-Guyatt scale the overall scientific quality of the Cochrane musculoskeletal reviews was good mean 5.02 (95% CI: 3.71 to 6.32).

Our study found that the reporting quality of Cochrane musculoskeletal systematic reviews was generally good, although there was room for improvement. For example, it might be feasible to develop specific guidelines for reporting protocols. Certainly more work is needed in reporting search results, documenting the flow of studies, identifying of the type of studies, and summarizing of the key findings.

Beverley Shea, Lex Bouter, Jeremy Grimshaw, Daniel Francis, Zulma Ortiz, George Wells, Peter Tugwell, and Maarten Boers. Scope for improvement in the reporting quality of systematic reviews from the Cochrane Musculoskeletal Group. J Rheumatol 2006 Jan;33(1):9-15.

Introduction

It has been estimated that a physician would need to read 17-20 journal articles a day to stay apprised of all research relevant to a particular area of clinical practice.[1] Since this is clearly impractical, interest has been growing in the use of pre-appraised and synthesized evidence resources such as systematic reviews, meta-analyses, and evidence-based clinical practice guidelines as aids to clinical decision-making.

Recognition of the need for systematic reviews of healthcare studies continues to grow and is indicated by the number of articles and empirical studies dealing with methods used in reviews, the number of systematic reviews published in healthcare journals, and the rapid growth of the Cochrane Collaboration.[2-4] The Cochrane Library is a compilation of systematic reviews designed to provide high quality scientific evidence on the effectiveness of various healthcare interventions.[5] Cochrane reviews serve an invaluable function by summarizing healthcare literature and supporting decisions for more effective clinical practices. The Cochrane Musculoskeletal Group (CMSG) is one of more than 50 entities within the Cochrane Collaboration. Its members are dedicated to preparing and maintaining systematic reviews of musculoskeletal conditions. Many interventions for gout, lupus erythematosus, osteoarthritis, osteoporosis, rheumatoid arthritis, soft tissue conditions, spondyloarthropathy, systemic sclerosis, and vasculitis have been reviewed. Separate review groups have studied reviews of back and musculoskeletal injuries. We assessed the quality of reviews conducted by the CMSG.

Growing recognition of the key role of reviews in synthesizing and disseminating research results has prompted careful scrutiny of the validity of reviews. In the 1970s and early 1980s, psychologists and social scientists drew attention to the systematic steps needed to minimize bias and random errors in literature reviews.[6-9] In the late 1980s, attention began to focus on the poor scientific quality of healthcare review articles.[10, 11]

An appreciation of the quality of a systematic review is essential to assess whether its recommendation of the use or avoidance of an intervention should be followed.[12, 13] Two major areas are assessed in determining the quality of a systematic review. The first is its methodological quality, which is an assessment of how well the systematic review was conducted (literature search, pooling of data, etc.). The second is its reporting quality, which is an assessment of how well its systematic reviewers have reported their methodology and findings. Separate tools were used to assess each quality area to obtain a comprehensive quality assessment. Assessors were sensitive to the fact that although methodological quality and reporting quality are intrinsically linked, a review may be strong in one area and weak in the other. It was also recognized that poor reporting makes it difficult to assess the methodological quality of a review.

Our objective was to review both the methodological and reporting quality of all published Cochrane systematic musculoskeletal reviews. This review should serve as a baseline, enabling the CMSG to measure improvement in both the methodological and the reporting quality of its reviews over time.

Materials and Methods

Choosing the Assessment Instruments

A review of published scales and checklists was performed to inventory the instruments available to assess the quality of systematic reviews (Chapter 2). Each item of a scale is scored numerically and individual numerical scores are combined to generate an overall quality score. To be considered a scale, an instrument should be able to measure across a continuum. On the other hand, a checklist provides an estimate of the overall quality of a review by using itemized criteria to assess individual aspects of reviews and facilitate their qualitative comparisons. Individual checklist items do not have numerical scores attached to them.

A literature search was conducted using Medline from January 1966 to February 1999 to identify all quality assessment instruments.[14, Chapter 2] three independent searches were completed using the following keywords: metaanalysis, review literature, systematic or quantitative or methodologic review, overview, review, information synthesis, integrative research review, guideline, checklist, tool, scoring, scale, clinimetric, quality, critical reading, methodology, and Medline. The "related articles" function was also used.

Twenty-six assessment instruments were found, including 23 checklists and three scales. The two instruments selected were a methodological assessment tool that had been rigorously developed by Oxman and Guyatt, known as the overview quality assessment questionnaire (OQAQ)[15, 16] and subsequently enhanced (see Chapter 2) and the reporting quality tool (Reporting quality of Meta-analysis (QUOROM)) developed by consensus.[17]

The overview quality assessment questionnaire

The specified purpose of the OQAQ[15] is to evaluate the scientific quality (i.e. adherence to scientific principles) of systematic reviews published in medical literature. It is not designed to measure literary quality, importance, relevance, originality, or any other esthetic or philosophical attribute of reviews.

The scale is divided into nine areas (Table 1). Question 10, the final question of the scale, requires each assessor to rate the overall scientific quality of each report. Possible scores on question 10 range from 1 to 7. This global question is answered based on how well the review scored on the first nine questions.[16]

Table 1
Scores for Cochrane musculoskeletal group (CMSG) systematic reviews from the overview quality assessment questionnaire (OQAQ)

Item	Questions	Yes n (%)	Partially or can't tell n (%)	No n (%)
1	Were the search methods used to find evidence reported?	50 (88)	3 (5)	4 (7)
2	Was the search strategy for evidence reasonably comprehensive?	36 (63)	3 (5)	18 (32)
3	Were the criteria used for deciding which studies to include in the overview reported?	57 (100)	0 (0)	0 (0)
4	Was bias in the selection of studies avoided?	30 (53)	24 (40)	4 (7)
5	Were the criteria used for assessing the validity of the included studies reported?	55 (97)	2 (3)	0 (0)
6	Was the validity of all the studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?	41 (72)	4 (7)	12 (21)
7	Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?	54 (95)	1 (2)	2 (3.5)
8	Were the findings of the relevant studies combined appropriately relative to the primary question the overview addressed?	57 (100)	0 (0)	0 (0)
9	Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?	55 (97)	1 (2)	1 (2)
10	How would you rate the scientific quality of this overview?	5.02 (95% CI 3.7 to 6.32)		

The Reporting quality of Meta-analysis (QUOROM) Checklist

The only tool found that was designed to assess the reporting quality of systematic reviews was the Reporting quality of Meta-analysis[17] statement consisting of a checklist and flow diagram. This checklist of standards for the reporting of meta-analyses describes the preferred way to present the Abstract, Introduction, Methods, Results, and Discussion sections of a report of meta-analysis. The checklist includes 18 items (Instrument 1). It requires authors of reviews to include a flow diagram that provides information about the number of studies identified, included, excluded, and the reasons for excluding them.

Table 2
Scores for Cochrane Musculoskeletal Group (CMSG) systematic reviews from Reporting quality of Meta-analysis (QUOROM)

Item	Questions	Yes n (%)	No n (%)
1	Does the title identify the report as a meta-analysis (or systematic review) of randomized trials?	3 (5)	54 (95)
2	Is the abstract in a structured format?	56 (98)	1 (2)
3	Do the objectives describe the clinical question explicitly?	56 (98)	1 (2)
4	Are the databases (i.e. list) and information sources described?	55 (96.5)	2 (3.5)
5	Is the selection criteria (population, intervention, outcome, and study design), methods for validity assessment, data abstraction, study characteristics, and quantitative data synthesis described in sufficient detail to permit replication?	47 (82)	10 (18)
6	Is there a description of the main results?	49 (86)	8 (14)
7	Are the conclusions presented?	57 (100)	0 (0)
8	Is the clinical, biologic rationale for the intervention and rationale for the review provided?	54 (95)	3 (5)
9	Were the information sources in detail (e.g., databases, registers, personal files, expert informants, agencies, hand-searching), and any restrictions (years considered, publication status, language of publication) provided?	52 (91)	5 (9)
10	Were the inclusion and exclusion criteria (defining population, intervention principal outcomes, and study design) presented?	57 (100)	0 (0)
11	Were the criteria and process used (e.g., masked conditions, quality assessment and their findings) in the validity assessment reported?	57 (100)	0 (0)
12	Was the process provided (e.g., completed independently, in duplicate)?	50 (88)	7 (12)
13	Was the type of study design, participants' characteristics, details of intervention, outcome definitions, and how clinical heterogeneity was assessed reported?	56 (98)	1 (2)
14	Were the principal measures of effect (e.g., relative risk), method of combining results (statistical testing and confidence intervals), handling of missing data, how statistical heterogeneity was assessed, a rationale for any a priori sensitivity and subgroup analyses, and any assessment of publication bias reported?	53 (93)	4 (7)
15	Was a meta-analysis profile summarizing trial flow provided?	0 (0)	57 (100)
16	Were the descriptive data for each trial presented?	57 (100)	0 (0)

17	Was the agreement on the selection and validity assessment reported? Were the simple summary results (for each treatment group in each trial, for each primary outcome) and data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (e.g., 2 x 2 tables of counts, means and standard deviations, proportions) reported?	56 (98)	1 (2)
18	Was a summarization of the key findings, discussion of clinical inferences based on internal and external validity, interpretation of the results in light of the totality of available evidence, description of potential biases in the review process (e.g., publication bias), and suggestion of a future research agenda presented?	47 (82)	10 (18)

The OQAQ and the QUOROM statement were found to be the instruments of choice to assess the quality of systematic reviews.

Literature

The study examined all published CMSG systematic reviews in the Cochrane Database of Systematic Reviews of the Cochrane Library, Issue 4, 2002.[5]

Pilot testing

Two assessors (DF, BS) conducted a pilot study to test the two quality assessment tools. As recommended by Glass et al.[7], agreement among assessors was maximized through consensus training involving discussion among reviewers. All inconsistencies identified were discussed and resolved during weekly meetings, and assessments were revised after consensus was reached. The pilot exercise was conducted to achieve a high level of inter-rater agreement. It was decided that an interclass correlation coefficient (ICC) of 0.60 would be the lowest acceptable level of inter-rater reliability. An ICC was calculated for the OQAQ, and the ICC results > 0.60 were eventually obtained, though initially there was less agreement between reviewers on the questions on the OQAQ. Therefore, the items were reviewed and discussed in detail, with agreement being reached on its scoring in cases where there had been disagreement. This pilot was repeated until an ICC > 0.60 was achieved. A final ICC of 0.81 (95% CI: 0.74 to 0.89) was obtained for the OQAQ. Minor refinements were incorporated into the interpretation of the scale, permitting a greater degree of precision when performing assessments. A kappa test was also performed on the CMSG reviews. The kappa was 0.321 (95% CI: 0.136 to 0.498) prior to discussion and consensus. Additional pilot testing improved the level of agreement slightly to 0.420 (95% CI: 0.322 to 0.514).

Data analysis

Data were extracted using prepared forms that included all the items of both quality instruments. All of the included reviews were assessed using this structured format. Frequencies were provided on all items of both instruments using SPSS 12.0. Percentages were recorded along with an overall mean score and confidence intervals.

Main study

The quality of the 57 included systematic reviews[18-74] was assessed by two independent reviewers using the two selected tools, with a third reviewer available when needed to reach a consensus (GW). One

reviewer (DF) had been involved with the Cochrane Collaboration for two and a half years and continues to work as a reviewer with the Acute Respiratory Infections Review Group in Brisbane, Australia. The second reviewer (BS) had been involved with the CMSG over the previous 10 years and was a co-author of some of the reviews.

Results

Using the OQAQ, the mean overall scientific quality of the 57 Cochrane musculoskeletal reviews evaluated was (mean 5.02 (95% CI: 3.71 to 6.32)). The scores for each item were of similar quality (Table 1). Of the 10 items making up the scale, the Cochrane musculoskeletal systematic reviews scored poorly on items 2 and 4: 63% of CMSG systematic reviews reported whether the search strategy for the evidence was reasonably comprehensive (item 2) and 53% reported whether study selection bias was avoided (item 4). All CMSG systematic reviews reported the criteria used for deciding which studies to include in the overview (item 3) and combined the findings of the relevant studies appropriately relative to the primary question addressed (item 8). The methods used for combining studies were reported in 95% of the reviews (item 7), while 97% of CMSG reviews reported the criteria used to assess the validity of included studies (item 5) and drew conclusions that were supported by the data and/or analysis reported in the overview (item 9). Eighty-eight percent of the reviews reported the search methods used to find evidence (item 1). However, only 72% reported whether the validity of all the studies referred to in the text was assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited) (item 6).

Scores on individual QUOROM items ranged from 5.0% (item 1) to 100% (items 7, 10, 11, 16) (Table 2). Only 5% of CMSG reviews identified the review as a meta-analysis or systematic review of randomized trials in their title (item 1). Almost all CMSG reviews had an abstract with a structured format (item 2) and included objectives (item 3) and data sources (item 4) in the abstract. Items on which they were less likely to report adequately were results (item 6) and selection criteria (item 5) (i.e. population, intervention, outcome, study design, methods for validity assessment, data abstraction, study characteristics, and quantitative data synthesis).

Almost 90% of the reviews described their method of data abstraction (item 12), while no review provided a flow chart for the included and excluded studies (item 15).

More than half the CMSG reviews received a rating of "adequate" on 50% or more of the 10 OQAQ quality items. On the overall quality item (range 0-7) (item 10), the CMSG reviews scored relatively well, with only minor flaws identified (mean 5.02 (95% CI: 3.71 to 6.32)) (Table 1). Of the 18 QUOROM items, the CMSG systematic reviews scored more than 50% on all but two of the items (items 1 and 15).

One item (18) was noted as being more difficult to assess and certainly needs further exploration (i.e., the summarization of the key findings, discussion of clinical inferences based on internal and external validity, interpretation of the results in light of the totality of available evidence, description of potential biases in the review process (e.g., publication bias), and suggestion of a future research agenda presented).

Documenting the flow of included and excluded studies and summarizing the results are two areas needing improvement in reporting.

Discussion

Assessments made using both quality instruments indicated that the quality of Cochrane musculoskeletal systematic reviews was good, although minor flaws were observed. This is important to users of CMSG reviews as it provides assurance that their results are relatively reliable. Although their methodological quality and reporting quality were found to be fair to good, there is room for improvement. For example,

it might be feasible to develop specific guidelines for reporting protocols, improving reporting search results, and documenting the flow of studies.

The quality of systematic reviews requires examination in order to substantiate the claim that they are the best evidence available to clinicians, health policymakers, and consumers.

The use of assessment tools to structure peer review systems can encourage quality improvement in systematic reviews. The Cochrane Collaboration has begun to achieve this objective through continual peer review of protocols, reviews, and updated reviews from the analytical process through to the report. The use of evidence-based criteria such as the QUOROM statement can contribute to the improvement of reporting quality over time by establishing consistent guidelines for the methodological quality of systematic reviews. At least two studies have addressed improvement or lack of improvement over time. A review of 86 English-language metaanalyses assessed every report on 14 items from six content areas believed to be critical in the methodological quality and reporting quality of metaanalyses. These items included study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. They found that only 24 of the 86 (28%) metaanalyses addressed all six content areas.[75] This survey was updated in 1992 with little change in the results.[76] A similar study by the authors of this paper showed that the quality of systematic reviews does improve over time, but that the differences on specific items remain variable.[77] A comparison of Cochrane versus paper-based reviews revealed similar results.[78]

Inadequate reporting[79-82] is a significant impediment to the assessment of the quality of systematic reviews. Essential criteria may be met in a given study without being adequately reported in a review. In such a case, a study of high quality may appear to be poor in a review. It may be inaccurate to assume that items not included in a systematic review were missing in the study reviewed, but this is what users of systematic reviews are likely to do. Assessors of a systematic review may invest the time and effort required to obtain additional data directly from the investigators who conducted the systematic review, but ordinary readers cannot reasonably be expected to do so.

The ongoing use of reporting quality checklists[6, 80-82] is to be encouraged as it will facilitate the assessment of study validity by ensuring a high level of congruence between the quality of individual studies and their depiction in systematic reviews. Such an improvement in the reporting quality of reviewed studies will serve to make systematic reviews more accurate, reliable, persuasive, and useful to those who depend on them.

We tried to address the issue of potential conflict of interest by inviting someone from outside the Musculoskeletal Group to work with the team on a voluntary basis to carry out this study, a common practice for Cochrane work, especially methods work. One of the main reviewers of the studies is from Brisbane, Australia, and had worked with the Respiratory Airways group for approximately three years and is very well aware of the format of Cochrane reviews. The authors felt that because he was not involved in any of the CMSG reviews he would be a non-biased reviewer. We believed this would negate any potential biases.

Although the two instruments used in this study proved useful, the challenges that had to be overcome in applying them clearly demonstrated the need for better measurement instruments. Both instruments proved to be difficult to apply. The main problems encountered were lack of published guidance on their application and difficulties applying the overall global score. It was only after three rounds of pilot testing and resolution of several questions that arose concerning the application of the OQAQ assessment tool that the assessors were ready to apply it. Moreover, while using this scale, reviewers continued to encounter difficulties with its application. Although the QUOROM checklist was rather long and

time-consuming to apply, it was accompanied by more detailed directions regarding its use.

One additional item that was not addressed adequately due to the measurement tool not addressing it well was the inferences based on the results of the systematic reviews. More work is needed in this area.

The statistical analysis revealed that the reliability was poor to fair. There may be several explanations for this. First, the relative magnitude of the kappa value (i.e., the proportion of agreement beyond that expected by chance alone) is difficult to interpret. In the pilot test, the kappa values were categorized and labeled as suggested by Fleiss[83], but this classification is purely arbitrary. Kappa coefficient is a popular measure for chance-corrected nominal scale agreement between two raters. Future methodological work must include alternative options for calculating agreement among raters, such as exploring Bayesian inferences.

This study found that the overall quality of reports of Cochrane musculoskeletal systematic reviews was generally good, although, there was room for improvement. Particular areas that need special attention include the title and protocol, documentation of the flow of the studies, and inferences based on conclusions. A recent study reported that the methods for assessment of methodological quality of systematic reviews are still in their infancy and there is substantial room for improvement.[84]

Acknowledgements

The authors express their appreciation to David Moher for his helpful suggestions and to Ashley Porter and Thelma Hasson for their comments on the manuscript.

References

1. Davidoff F, Case K, Fried PW. Evidence-based medicine: why all the fuss? (editorial). *Ann Intern Med* 1995 May; 122(9): 727.
2. The Cochrane Methodology Register (CMR). The Cochrane Library. Chichester, UK: John Wiley & Sons Ltd; 2004: 3.
3. Yusuf S. Obtaining medically meaningful answers from an overview of randomized clinical trials. *Stat Med* 1987 Apr-May; 6(3): 281-94.
4. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988 Apr; 138(8): 697-703.
5. *The Cochrane Library.* Chichester, UK: John Wiley & Sons, Ltd.; 2003: 4.
6. Light RJ, Pillemer DB. A checklist for evaluating reviews. In: The systematic review. Cambridge, London: Harvard University Press 1984: 160-86.
7. Glass GV, McGraw B, Smith ML. Meta-analysis in social research Beverly Hills, CA: Sage Publications; 1981.
8. Rosenthal R. Meta-analytic procedures for social research. Rev. ed. Newbury Park, CA: Sage Publications; 1991.
9. Hunter JE, Schmidt FL, Jackson GB. Meta-analysis: Cumulating research findings across studies. In: Cooper HM, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994.
10. Mulrow CD. The medical review article: state of the science. *Ann Intern Med* 1987 Mar; 106(3): 485-88.
11. National Health Service (NHS) Centre for Reviews and Dissemination (CRD). York, UK: University of York; Y010 5DD.
12. Moher D, Jadad A, Nichol G, Penman M, Tugwell T, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Contr Clin Trials* 1995 Feb; 16(1): 62-73.
13. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998 Jul; 280(3): 278-80.
14. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. Systematic review in health care meta-analysis in context. London: BMJ Books 2001: 122-39.
15. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44(11): 1271-78.
16. Oxman AD. Checklists for review articles. *BMJ* 1994; 309: 648-51.
17. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Reporting quality of Meta-analyses. *Lancet* 1999; 354: 1896-900.
18. Blumenauer B, Judd M, Wells G, et al. Infliximab for the treatment of rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
19. Bonaiuti D, Shea B, Iovine R, et al. Exercise for preventing and treating osteoporosis in postmenopausal women. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
20. Brosseau L, Casimiro L, Milne S, et al. Deep transverse friction massage for treating tendonitis. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.
21. Brosseau L, Welch V, Wells G, et al. Low level laser therapy (classes I, II and III) for treating osteoarthritis. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.
22. Brosseau L, Welch V, Wells G, et al. Low level laser therapy (classes I, II and III) for treating rheumatoid arthritis. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.

23. Brosseau L, Casimiro L, Robinson V, et al. Therapeutic ultrasound for treating patellofemoral pain syndrome. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
24. Buchbinder R, Green S, White M, Barnsley L, Smidt N, Assendelft WJJ. Shock wave therapy for lateral elbow pain. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.
25. Buchbinder R, Green S, Bell S, Barnsley L, Smidt N, Assendelft WJJ. Surgery for lateral elbow pain. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.
26. Busch A, Schachter CL, Peloso PM, Bombardier C. Exercise for treating fibromyalgia syndrome. In: *The Cochrane Library,*, Issue 4, 2002. Oxford: Update Software.
27. Casimiro L, Brosseau L, Milne S, Robinson V, Wells G, Tugwell P. Acupuncture and electroacupuncture for the treatment of RA. In: The Cochrane Library, Issue 4, 2002. Oxford: Update Software.
28. Casimiro L, Brosseau L, Robinson V, et al. Therapeutic ultrasound for the treatment of rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
29. Clark P, Tugwell P, Bennet K, et al. Injectable gold for rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
30. Cranney A, Welch V, Adachi JD, et al. Calcitonin for preventing and treating corticosteroid-induced osteoporosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
31. Cranney A, Welch V, Adachi JD, et al. Etidronate for treating and preventing postmenopausal osteoporosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
32. Criswell LA, Saag KG, Sems KM, et al. Moderate-term, low-dose corticosteroids for rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
33. Dagfinrud H, Hagen K. Physiotherapy interventions for ankylosing spondylitis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
34. Garner S, Fidan D, Frankish R, et al. Celecoxib for rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
35. Garner S, Fidan D, Frankish R, et al. Rofecoxib for rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
36. Gotzsche PC, Johansen HK. Short-term low-dose corticosteroids vs placebo and nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. In: The Cochrane Library, Issue 4, 2002. Oxford: Update Software.
37. Green S, Buchbinder R, Barnsley L, et al. Acupuncture for lateral elbow pain. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
38. Green S, Buchbinder R, Glazier R, Forbes A. Interventions for shoulder pain. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
39. Green S, Buchbinder R, Barnsley L, et al. Non-steroidal anti-inflammatory drugs (NSAIDs) for treating lateral elbow pain in adults. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
40. Haguenauer D, Welch V, Shea B, Tugwell P, Wells G. Fluoride for treating postmenopausal osteoporosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
41. Homik J, Cranney A, Shea B, et al. Bisphosphonates for steroid induced osteoporosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
42. Homik J, Suarez-Almazor ME, Shea B, Cranney A, Wells G, Tugwell P. Calcium and vitamin D for corticosteroid-induced osteoporosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
43. Hulme J, Robinson V, de Bie R, Wells G, Judd M, Tugwell P. Electromagnetic fields for the treatment of osteoarthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
44. Jones G, Crotty M, Brooks P. Interventions for treating psoriatic arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
45. Karjalainen K, Malmivaara A, van Tulder M, et al. Multidisciplinary rehabilitation for fibromyalgia and musculoskeletal pain in working adults. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.

46. Little CV, Parsons T, Logan S. Herbal therapy for treating osteoarthritis (Cochrane Review). In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software; 2002.
47. Little C, Parsons T. Herbal therapy for treating rheumatoid arthritis (Cochrane Review). In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software; 2002.
48. Ortiz Z, Shea B, Suarez-Almazor M, Moher D, Wells G, Tugwell P. Folic acid and folinic acid for reducing side effects in patients receiving methotrexate for rheumatoid arthritis (Cochrane Review). In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software; 2002.
49. Osiri M, Welch V, Brosseau L, et al. Transcutaneous electrical nerve stimulation for knee osteoarthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
50. Pelland L, Brosseau L, Casimiro L, Robinson V, Tugwell P, Wells G. Electrical stimulation for the treatment of rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
51. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman D. Cyclofenil for Raynaud's phenomenon in progressive systemic sclerosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
52. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman D. Iloprost and cisaprost for Raynaud's phenomenon in progressive systemic sclerosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
53. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman D. Ketanserin for Raynaud's phenomenon in progressive systemic sclerosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
54. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman D. Prazosin for Raynaud's phenomenon in progressive systemic sclerosis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
55. Riemsma RP, Kirwan JR, Taal E, Rasker JJ. Patient education for adults with rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
56. Robinson V, Brosseau L, Casimiro L, et al. Thermotherapy for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
57. Saenz A, Ausejo M, Shea B, Wells G, Welch V, Tugwell P. Pharmacotherapy for Behcet's syndrome. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
58. Struijs PAA, Smidt N, Arola H, van Dijk CN, Buchbinder R, Assendelft WJJ. Orthotic devices for the treatment of tennis elbow. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
59. Suarez-Almazor ME, Belseck E, Shea B, Homik J, Wells G, Tugwell P. Antimalarials for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
60. Suarez-Almazor ME, Spooner C, Belseck E, Shea B. Auranofin versus placebo in rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
61. Suarez-Almazor ME, Spooner C, Belseck E. Azathioprine for treating rheumatoid arthritis (Cochrane Review). In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software; 2002.
62. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Cyclophosphamide for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
63. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Methotrexate for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
64. Suarez-Almazor ME, Spooner C, Belseck E. Penicillamine for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
65. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Sulfasalazine for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
66. Takken T, van der Net J, Helders PJM. Methotrexate for treating juvenile idiopathic arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
67. Towheed T, Shea B, Wells G, Hochberg M. Analgesia and non-aspirin, non-steroidal anti-inflammatory drugs for osteoarthritis of the hip. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.

68. Towheed TE, Anastassiades TP, Shea B, Houpt J, Welch V, Hochberg MC. Glucosamine therapy for treating osteoarthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
69. Trevisani VFM, Castro AA, Neves Neto JF, Atallah AN. Cyclophosphamide versus methylprednisolone for treating neuropsychiatric involvement in systemic lupus erythematosus. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
70. van den Ende CHM, Vliet Vlieland TPM, Munneke M, Hazes JMW. Dynamic exercise therapy for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
71. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy for rheumatoid arthritis and osteoarthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
72. Watson MC, Brookes ST, Kirwan JR, Faulkner A. Non-aspirin, non-steroidal anti-inflammatory drugs for treating osteoarthritis of the knee. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
73. Welch V, Brosseau L, Peterson J, Shea B, Tugwell P, Wells G. Therapeutic ultrasound for osteoarthritis of the knee. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
74. Wells G, Haguenauer D, Shea B, Suarez-Almazor ME, Welch VA, Tugwell P. Cyclosporine for treating rheumatoid arthritis. In: *The Cochrane Library,* Issue 4, 2002. Oxford: Update Software.
75. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987; 316: 450-55.
76. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. *Mt Sinai J Med* 1996; 3-4: 216-24.
77. Shea B. Assessing the reporting quality meta-analyses of randomized controlled trials. MSc thesis. Ottawa: University of Ottawa; 1999.
78. Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the reporting quality of Cochrane reviews and systematic reviews published in paper-based journals. *Eval Health Prof* 2002; 25: 116-29.
79. The Standards of Reporting Trials Group (SORT). A proposal for structured reporting of randomized controlled trials. *JAMA* 1994; 272: 1926-31.
80. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; 272: 125-28.
81. Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature. *Ann Intern Med* 1994; 121: 894-95.
82. Begg C, Cho M, Eastwood S, et al. Improving the reporting quality of randomized controlled trials. The CONSORT statement. *JAMA* 1996; 276: 637-39.
83. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981.
84. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. *BMJ* 2005; 330: 1053. Epub 2005 Apr 7.

4

4

DOES UPDATING IMPROVE THE METHODOLOGICAL AND REPORTING QUALITY OF REVIEW QUALITY AND THE REPORTING QUALITY OF COCHRANE REVIEWS?

Systematic reviews are of variable methodological quality. Updating a systematic review provides an opportunity to improve its quality. This study assessed the methodological and reporting quality of updated Cochrane systematic reviews.

We identifed updated Cochrane systematic reviews published in Issue 4, 2002 of the Cochrane Library. We assessed the updated and original versions of the systematic reviews using two instruments: the 10-item overview quality assessment questionnaire (OQAQ), and an 18-item reporting quality checklist and flow chart based upon the Reporting quality of Meta-analyses (QUOROM) statement. For each included systematic review, at least two reviewers extracted data and assessed quality. We calculated the percentage (with a 95% confidence interval) of 'yes' answers to each question. We calculated mean differences in percentage, 95% confidence intervals, and p-values for each of the individual items and the overall methodological quality score of the updated and pre-updated versions using OQAQ.

In total, we assessed 53 systematic reviews. There was no significant improvement with updating in the global quality score of the OQAQ (mean difference 0.11 (95% CI: -0.28 to 0.70 p=0.52)). Updated reviews showed a significant improvement of 18.9% (95% CI: 7.2 to 30.6 p<.01) on the OQAQ item assessing whether the conclusions drawn by the author(s) were supported by the data and /or analysis presented in the systematic review. There is clearly room for improvement of methodological quality. The QUOROM statement showed that the reporting quality of Cochrane reviews improved in some areas with updating, but leaves room for further improvement. These most notable improvements were seen on the items relating to data sources with a significant difference of 17.0% (95% CI: 9.8 to 28.7 p=0.01) review methods 35% (95% CI: 24.1 to 49.1 p=0.00), searching methods 18.9% (95% CI: 9.7 to 31.6 p=0.01), and data abstraction 18.9% (95% CI: 11.7 to 30.9 p=0.00).

Updated Cochrane reviews appeared to improve on some aspects of reporting and methodological quality. There is clearly room for improvement of overall methodological quality. Authors updating reviews should address identified methodological or reporting weaknesses.

Beverley Shea, Maarten Boers, Jeremy Grimshaw, Candyce Hamel, Lex Bouter. Does updating improve the methodological and reporting quality of review quality and the reporting quality of Cochrane reviews? *BMC Medical Research Methodology* 2006; 6: 27.

Background

A number of papers have been published on the methodological and reporting quality of reviews. A review of 86 English language meta-analyses published between 1950 and 1986 by Sacks[1] assessed every report on 14 items covering six content areas believed to be critical in the reporting of meta-analysis. Only 28% of these meta-analyses were found to address all six content areas. An updated survey in 1992 showed little change.[2] Shea[3] compared the methodological quality of paper-based and electronic systematic reviews and found little difference and a lot of room for improvement. Assendelft[4] reviewed 51 reviews and noted that reviews that favoured a given intervention tended to have higher methodological quality scores. In contrast, Jadad and McQuay[5] reviewed 80 systematic reviews published between 1980 and 1992 and found a disconcerting link amongst reviews where results favoured an intervention and reviews of poor methodological quality. Jadad found that Cochrane reviews had greater methodological rigor, more frequent updates, and higher overall quality scores than those published in peer-reviewed paper journals, although, both types were found to contain extensive and serious flaws.

The Cochrane Collaboration (www.Cochrane.org) is an international not-for-profit organization that conducts and updates systematic reviews of healthcare studies. With the exception of a few studies completed by Jadad[6, 7] and Shea[8], little is known about the quality of Cochrane reviews and whether their quality is improving over time and with updating. To our knowledge, the impact of updating on their quality has neither been evaluated nor studied.

Studies of the quality of systematic reviews can focus on methodological or reporting quality. Methodological quality is concerned with how well a systematic review was designed and conducted (e.g. literature searching, pooling of data, etc.). Reporting quality considers how well systematic reviewers have reported their methodology and findings.

The purpose of our study was to compare the methodological and reporting quality of Cochrane systematic reviews with that of their updated versions in order to determine whether updating contributed significantly to the improvement of their quality in these two dimensions.

Methods

We selected all updated Cochrane systematic reviews from the Cochrane Database of Systematic Reviews 2002[9] and the same reviews prior to their update (Annex 1). Updated reviews were chosen following the definition of 'updating' included in the Cochrane handbook.[10]

Based on a previously published study[11], the two instruments chosen to assess the quality of Cochrane systematic reviews for this study were the enhanced overview quality assessment questionnaire (OQAQ)[11-13] (Instrument 2) and the Reporting quality of Meta-analysis (QUOROM) checklist (Instrument 1).[14]

The OQAQ was selected because it had strong face validity, provided data on several essential elements of its development, and had available a published assessment of its construct validity.[12] In addition, its validity had been thoroughly tested and established using a number of independent measures.[13] However, we noted difficulty applying its questions so we developed an enhanced version of the OQAQ which incorporated guidelines for their use.[11] The OQAQ scale measures across a continuum using nine questions (items 1-9) designed to assess various aspects of the methodological quality of systematic reviews and one overall assessment question (item 10). When the scale is applied to a systematic review, the first nine items are scored by selecting either yes, no, partial/can't tell. The tenth item requires assessors to assign an overall quality score on a seven-point scale.[12]

The QUOROM statement was chosen to assess reporting quality. Although this checklist has not yet been fully validated, extensive work on it has been conducted and reported.[14] The QUOROM statement is comprised of a checklist and flow diagram (http://www.consort-statement.org/QUOROM.pdf) and was developed using a consensus process designed to strengthen the reliability of the estimates it yields when applied by different assessors. It estimates the overall reporting quality of systematic reviews. The checklist questions whether authors have provided readers with information on 18 items, including searches, selection, validity assessment, data abstraction, study characteristics, quantitative data syntheses, and trial flow. It also questions whether authors have included a flow diagram with information about the number of randomized controlled trials identified, included and excluded, and the reasons for any exclusion. Individual checklist items included in this instrument are answered either yes, no or partial/can't tell.

For each included Cochrane systematic review, we calculated the percentage (with a 95% confidence interval) of 'yes' answers to each item of OQAQ and QUOROM. Individual Cochrane systematic reviews were then compared to their updated versions with respect to the percentage of 'yes' ratings received. A percentage difference (with a 95% confidence interval and p-values) of 'yes' answers was calculated for each individual question. In addition, a mean difference (with a 95% confidence interval and p-value) was calculated for the overall quality score (OQAQ) of the updated and pre-updated versions.

We assessed the practicability of the OQAQ instrument by recording the time it took to complete scoring and the instances where scoring was difficult. We obtained data on clarity, ambiguity, completeness, and user-friendliness.

Assessments of all individual reviews were conducted independently by at least two reviewers. One of these reviewers (CH) had been involved with the Cochrane Collaboration for six years. The other reviewer (BS) had been involved with the Cochrane Collaboration for ten years and had carried out a number of systematic reviews. A third reviewer (DF) was available to assist in the resolution of assessment discrepancies.

Results

In total, we included 53 reviews. The mean period between the publication of the original and updated versions of the reviews was 2.7 years (range four months to five years). Complete results for both instruments applied to the reviews are provided in Tables 1 and 2, but highlights are summarized below.

OQAQ

Table 1 presents the methodological quality assessments obtained using this scale. There was no significant difference in the global assessment (item 10 - How would you rate the scientific quality of the overview?) (mean score original review 4.70, mean score updated review 4.81, difference in means +0.11 (95% CI: -0.28 to 0.70 p=0.52)).

There were improvements on seven individual items, although only one item showed a significant improvement (item 9 - Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?) (percent original reviews complying with item 76% (95% CI: 4.47 to 4.94) percentage updated reviews complying with item 94% (95% CI: 88.1 to 100.0) difference +18.9% (95% CI: +7.2% to +30.6% p<0.01).

Assessors reported difficulties scoring item 10, specifically in the absence of a meta-analysis.

QUOROM checklist

Table two presents the assessments of reporting quality obtained using the Quality of Reviews of Meta-analysis (QUOROM) checklist. Scores (awarded for 'yes' response) on the 18 items for original Cochrane reviews ranged from 0% (item 1) to 100% (items 2 and 18). Four of the 18 items revealed notable improvements relating to data sources (item 4) with a significant difference of 17.0% (95% CI: 9.8 to 28.7 p=0.01), review methods (items 5) with a difference of 35% (95% CI: 24.1 to 49.1 p=0.00), searching methods (item 9) with a difference of 18.9% (95% CI: 9.7 to 31.6 p=0.01), and data abstraction (item 12) with a difference of 18.9% (95% CI: 11.7 to 30.9 p=0.00).

Three questions, item (10) -3.7% (95% CI: -9.0 to 3.8 p=0.15), item (14) -3.7% (95% CI: -14.8 to 6.4 p=0.62) and item (16), -5.7% (95% CI: -15.8 to 3.0 p=0.40) had lower mean scores on updated reviews than on original reviews, but these differences were not statistically significant. There is room for improvement of the reporting quality.

Assessors reported difficulty with overlap and the length of the questionnaire.

Discussion

In assessing the quality of the sample of 53 systematic reviews of the Cochrane Collaboration, two assessment tools were used. The larger improvement on the QUOROM checklist than on the OQAQ suggests that although the reporting quality has improved slightly, the quality of design and conduct has not changed. For example, the items in reporting of selection criteria and searching are improved on QUOROM, but the equivalent items in OQAQ which relate to how well these were carried out have not changed.

The significant improvements for the OQAQ item assessing whether the conclusions drawn by the author(s) were supported by the data and /or analysis reported in the overview is also worthy of note. This could suggest that authors are paying more attention to stating their conclusions in relation to the data provided. This item has been reported by other methodologists in the context of assessing quality.[15, 16]

Improvement in the reporting quality of the abstract, as assessed by the QUOROM instrument, might suggest that authors are paying more attention to this important area. Reporting of literature searches improved, however, the literature searches themselves did not improve. The increased involvement of library scientists could be further explored.

Assessors found the two instruments used in the study to have associated practical weaknesses. The results they generated permitted comparisons between original and updated reviews, but their combined length (28 items) made their use somewhat cumbersome and inefficient. Another problem encountered was the lack of adequate published guidance on the proper application of the OQAQ. Assessors continued to encounter difficulties in using the instruments. This may account for the low rating on question 10. Individually the items scored well but the overall general rating indicated flaws. The QUOROM checklist provided assistance with fairly detailed user instructions. It was found though, that the instructions were quite time-consuming to apply.[11] Another problem noted was that several of the questions asked by these two instruments appeared to cover the same subject matter, raising questions of face validity.

The methodological and reporting quality of Cochrane reviews was found to be reasonable, though further improvement is obviously needed in both areas. A recent study by Moja[15] drew the same conclusion. In addition, the quality-improvement impact of updating was found to be relatively minor.

On some assessed factors, particularly with respect to reporting quality, updated reviews actually scored lower than original reviews. Currently, Cochrane review updates are carried out primarily to incorporate new findings rather than to improve quality. It would be beneficial for updates to also address reporting issues such as the omission of methodological descriptions. As well, updates should attempt to improve the methodological weaknesses in reviews.

The Cochrane Collaboration endeavours to improve the quality of its systematic reviews through the application of a continuous peer review process during its development. The effectiveness of the peer-review process might be improved by increased attention to areas of reporting and/or methodological weaknesses. Reviewers should adhere more faithfully to the guidelines provided in the Cochrane handbook[16] in order to improve the methodological quality of reviews and to the QUOROM statement[14] to improve the reporting quality of systematic reviews.

Conclusion

Updated Cochrane reviews appeared to improve in some aspects of methodological quality and reporting quality. There is clearly room for improvement in overall methodological quality. Authors updating reviews should address identified methodological or reporting weaknesses.

Acknowledgements

The authors would like to acknowledge Daniel Francis for his assistance with quality assessment, Mike Clarke, Ron Habinski and Crystal Huntly-Ball for their comments on earlier drafts of this paper; Andy Oxman and David Moher for their previous work on similar methods; Tim Ramsay for his statistical advice, and our external reviewer Penny Whiting for her helpful suggestions.

Table 1
Comparisons of percent 'yes' and percent 'differences' using the QOAQ Scale

Questions OQAQ	Original Reviews % reviews yes (95% CI)	Updated Reviews % reviews yes (95% CI)	Difference % Yes (95% CI) p-value
1. Were the search methods used to find evidence reported?	81% (68.9 to 93.4)	87% (75.7 to 97.9)	5.7% (-6.6 to 17.9) 0.43
2. Was the search strategy for evidence reasonably comprehensive?	68% (47.0 to 88.7)	70% (48.3 to 91.3)	1.9% (-19.0 to 22.8) 0.83
3. Were the criteria used for deciding which studies to include in the overview reported?	98% (94.4 to 100.0)	91% (80.5 to 100.0)	-7.6% (-11.2 to 3.9) 0.09
4. Was bias in the selection of studies avoided?	64% (45.4 to 82.8)	70% (52.0 to 87.6)	5.7% (-13.1 to 24.4) 0.53
5. Were criteria used for assessing validity of the included studies reported?	89% (73.7 to 100.0)	96% (85.9 to 100.0)	7.6% (-7.4 to 22.5) 0.14
6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing studies that are cited)?	85% (70.6 to 99.2)	93% (82.9 to 100.0)	7.6% (-0.68 to 21.9) 0.22
7. Were methods used to combine the findings of relevant studies (to reach a conclusion) reported?	89% (66.2 to 99.8)	83% (71.5 to 100.0)	-5.7% (-22.9 to 11.6) 0.40
8. Were findings of the relevant studies combined appropriately relative to the primary question addressed?	87% (68.4 to 100.0)	91% (78.7 to 100.0)	3.8% (-14.6 to 22.2) 0.54
9. Were the conclusions made by the author(s) supported by the data and /or analysis reported in the overview?	76% (63.8 to 87.2)	94% (88.1-100.0)	18.9% (7.2 to 30.6) 0.01
10. How would you rate the scientific quality of this overview?	4.70 (4.47 to 4.94)	4.81 (4.59 to 5.04)	0.11 (-.28 to .70) 0.52

Table 2
Comparisons of percent 'yes' and percent 'differences' using the reporting quality of meta-analysis (QUOROM)

Questions QUOROM			Original	Updated	Difference
Items	**Heading** *Subheading*	**Descriptor**	**Reviews % reviews yes (95% CI)**	**Reviews % reviews yes (95% CI)**	**Difference % (95% CI) p-value**
1	**Title**	Identified the report as a meta-analysis (or systematic review) of randomized trials.	0,0%	0,0%	0% NS
2	**Abstract**	Used a structured format.	100%	100%	0% NS
3	*Objectives*	The clinical question explicitly.	96.2% (91.0 to 100.0)	98% (94.4 to 100.0)	1.9% (-1.8 to 7.0) 0.56
4	*Data sources*	The databases (e.g. list) and other information sources.	76% (63.8 to 87.2)	93% (85.3 to 99.6)	17.0% (9.8 to 28.7) 0.01
5	*Review methods*	The selection criteria (e.g. population, intervention, outcome, and study design; methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis) in sufficient detail to permit replication.	40% (26.3 to 52.9)	76% (63.8 to 87.2)	35.9% (24.1 to 49.1) 0.00
6	*Results*	Characteristics of the randomized trials included and excluded; qualitative and quantitative findings (e.g. point estimates and confidence intervals); and subgroup analyses.	40% (26.3 to 52.9)	76% (63.8 to 87.2)	35.9% (24.1 to 49.1) 0.00
7	*Conclusion*	The main results.	96% (91.4 to 100.0)	98% (94.4 to 100.0)	1.9% (-1.8 to 7.1) 0.56
8	*Intro*	The explicit clinical problem, biologic rationale for the intervention, and rationale for review.	98% (94.4 to 100.0)	100%	2% (1.9 to 5.6) 0.31
9	**Methods** *Searching*	The information sources, in detail (e.g., databases, registers, personal files, expert informants, agencies, hand-searching), and any restrictions (e.g. years considered, publication status, language of publication).	68% (55.2 to 80.6)	87% (77.7 to 96.0)	18.9% (9.7 to 31.6) 0.02

10	*Selection*	The inclusion and exclusion criteria (defining population, intervention principal outcomes, and study design).	100%	96% (91.0 to 100.0)	-3.7% (-9.0 to 3.8) 0.15
11	*Validity assessment*	The criteria and process used (e.g., masked conditions, quality assessment and their findings.	87% (77.6 to 96.0)	96% (91.0 to 100.0)	9.4% (4.3 to 18.6) 0.08
12	*Data abstraction*	The process used (e.g., completed independently, in duplicate).	74% (61.6 to 85.6)	93% (85.3 to 99.6)	18.9% (11.7 to 30.9) 0.01
13	*Study characteristics*	The type of study design, participants' characteristics, details of intervention, outcome definitions, etc.; and how clinical heterogeneity was assessed.	89% (61.6 to 85.6)	96% (91.0 to 100.0)	7.6% (2.4 to 16.1) 0.14
14	*Quantitative data synthesis*	The principal measures of effect (e.g., relative risk), method of combining results (statistical testing and confidence intervals), handling of missing data, etc.; how statistical heterogeneity was assessed; a rationale for any a priori sensitivity and subgroup analyses; and any assessment of publication bias.	83% (72.8 to 93.2)	79% (68.2 to 90.2)	-3.7% (-14.8 to 6.4) 0.62
15	**Results** *Trial flow*	Provide a meta-analysis profile summarizing trial flow.	2% (-1.8 to 5.6)	8% (0.36 to 14.7)	5.7% (-1.5 to 9.4) 0.17
16	*Study characteristics*	Present descriptive data for each trial (e.g., age, sample size, intervention, dose, and duration, follow-up).	89% (80.1 to 97.3)	83% (72.9 to 93.2)	-5.7% (-15.9 to 2.9) 0.40
17	*Quantitative data synthesis*	Report agreement on the selection and validity assessment; present simple summary results (for each treatment group in each trial, for each primary outcome); data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (e.g., 2 x 2 tables of counts, means and standard deviations, proportions).	81% (70.5 to 91.8)	89% (80.1 to 97.3)	7.5 (-1.1 to 18.2) 0.28
18	**Discussion**	Summarize the key findings; discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process (e.g., publication bias); and suggest a future research agenda.	96% (91.0 to 100.0)	96% (91.0 to 100.0)	0% (-5.2 to 5.2) NS

References

1. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987; 316: 450-55.
2. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. *Mt Sinai J Med* 1996; 63: 216-24.
3. Shea B. Assessing the reporting quality meta-analyses of randomized controlled trials. MSc thesis. University of Ottawa, Department of Epidemiology and Community Medicine; 1999.
4. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA 1995; 274: 1942-48.
5. Jadad AR, McQuay HJ. Meta-analyses to evaluate analogies interventions: a systematic qualitative review of their methodology. *J Clin Epidemiol* 1996; 49: 235-43.
6. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA 1998; 280: 278-80.
7. Jadad AR, Moher M, Brownman GP, Booker L, Sigouin C, Fuentes M, Stevens R. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. BMJ 2000; 320: 537-40.
8. Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the reporting quality of Cochrane review and systematic reviews published in paper-based journals. *Eval Health Prof* 2002 Mar; 25(1): 116-29.
9. The Cochrane Library, 2002 Issue 4, Chichester, UK: John Wiley & Sons, Ltd.
10. M Clarke, AD Oxman. Cochrane Reviewers Handbook. *The Cochrane Library*, Issue, 2002.
11. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In Systematic Reviews in Health Care: Meta-analyses in Context. 2nd edition. Edited by BMJ Publishing Group. London: Eager M et al 2001; 122-39. (Chapter 2)
12. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44: 1271-78.
13. Oxman AD. Checklists for review articles. BMJ 1994; 309: 648-51.
14. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Reporting quality of meta-analyses. *Lancet* 1999; 354: 1896-900.
15. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A and on behalf of the Metaquality Study Group. Assessment of methodological quality of primary studies by systematic review: results of the metaquality cross sectional study. *BMJ* 2005; 330: 1053-58; originally published online 7 Apr 2005; doi:10.1136/bmj.38414.515938.8F.
16. The Cochrane Library, Issue 4, 2005, Chichester, UK: John Wiley & Sons, Ltd.

Annex 1: List of updated systematic reviews

1. Parker MJ, Handoll HHG, Griffiths R. Anaesthesia for hip fracture surgery in adults (Updated). *The Cochrane Library* Issue 4; 2001.
2. Wark P, Wilson AW, Gibson PG. Azoles for allergic bronchopulmonary aspergillosis associated with asthma (Updated). *The Cochrane Library* Issue 4; 2001.
3. Bara AI, Barley EA. Caffeine for asthma (Updated). *The Cochrane Library* Issue 4; 2001.
4. Lima AR, Lima MS, Soares BGO, Farrell M. Carbamazepine for cocaine dependence (Updated). *The Cochrane Library* Issue 4; 2001.
5. McIntosh HM. Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria (Updated). *The Cochrane Library* Issue 4; 2001.
6. Henderson-Smart DJ, Subramanian P, Davis PG. Continuous positive airway pressure versus theophylline for apnea in preterm infants (Updated). *The Cochrane Library* 4; 2001.
7. Selak V, Farquhar C, Prentice A, Singla A. Danazol for pelvic pain associated with endometriosis (Updated). *The Cochrane Library* Issue 4; 2001.
8. Symington A, Pinelli J. Developmental care for promoting developments and preventing morbidity in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
9. Henderson-Smart DJ, Steer PA. Doxapram treatment for apnea in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
10. Mwandumba HC, Squire SB. Fully intermittent dosing with drugs for treating tuberculosis in adults (Updated). *The Cochrane Library* Issue 4; 2001.
11. Olin J, Schneider L. Galantamine for Alzheimer's disease (Updated). *The Cochrane Library* Issue 4; 2001.
12. Candelise L, Ciccone A. Gangliosides for acute ischaemic stroke (Updated). *The Cochrane Library* Issue 4; 2001.
13. Tubman TRJ, Thompson SW. Glutamine supplementation for preventing morbidity in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
14. Askie LM, Henderson-Smart DJ. Gradual versus abrupt discontinuation of oxygen in preterm or low birth weight infants (Updated). *The Cochrane Library* Issue 4; 2001.
15. Farquhar C, Vandekerckhove P, Arnot M, Lilford R. Laparoscopic "drilling" by diathermy or laser for ovulation induction in anovulatory polycystic ovary syndrome (Updated). *The Cochrane Library* Issue 4; 2001.
16. Simmer K. Longchain polyunsaturated fatty acid supplementation in infants born at term (Updated). *The Cochrane Library* Issue 4; 2001.
17. Counsell C, Sandercock P. Low-molecular-weight heparins or heparinoids versus standard unfractionated heparin for acute ischaemic stroke (Updated). *The Cochrane Library* Issue 4; 2001.
18. Villar J, Khan-Neelofur D. Patterns of routine antenatal care for low-risk pregnancy (Updated). *The Cochrane Library* Issue 4; 2001.
19. van der Velden J, Ansink A. Primary groin irradation vs primary groin surgery for early vulvar cancer (Updated). *The Cochrane Library* Issue 4; 2001.
20. Henderson-Smart DJ, Steer PA. Prophylactic caffeine to prevent postoperative apnea following general anesthesia in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
21. Gulmezoglu AM, Forna F, Villar J, Hofmeyr GJ. Prostaglandins for prevention of postpartum haemorrhage (Updated). *The Cochrane Library* Issue 4; 2001.
22. Askie LM, Henderson-Smart DJ. Resticted versus liberal oxygen exposure for preventing morbidity and mortality in preterm or low birth weight infants (Updated). *The Cochrane Library* Issue 4; 2001.
23. Miller RG, Mitchell JD, Moore DH. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND) (Updated). *The Cochrane Library* Issue 4; 2001.

24. van Pinxteren B, Numans ME, Bonis PA, Lau J. Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease (Updated). *The Cochrane Library* Issue 4; 2001.
25. Bradley P, Lindsay B. Specialist epilepsy nurses for treating epilepsy (Updated). *The Cochrane Library* Issue 4; 2001.
26. Farquhar C, Lee O, Toomath R, Jepson R. Spironolactone versus placebo or in combination with steroids for hirsutism and/or acne (Updated). *The Cochrane Library* Issue 4; 2001.
27. Scott HD, Laake K. Statins for the reduction of risk of Alzheimer's disease (Updated). *The Cochrane Library* Issue 4; 2001.
28. Stevens B, Ohlsson A. Sucrose for analgesia in newborn infants undergoing painful procedures (Updated). *The Cochrane Library* 4; 2001.
29. Carroli G, Bergel E. Umbilical vein injection for management of retained placenta (Updated). *The Cochrane Library* Issue 4; 2001.
30. Demicheli V, Rivetti D, Deeks JJ, Jefferson TO. Vaccines for preventing influenza in healthy adults (Updated). *The Cochrane Library* 4; 2001.
31. Soares KVS, McGrath JJ. Vitamin E for neuroleptic-induced tardive dyskinesia (Updated). *The Cochrane Library* Issue 4; 2001.
32. Sipe J, Dunn L. Aciclovir for Bell's palsy (idiopathic facial paralysis) (Updated). *The Cochrane Library* Issue 4; 2001.
33. Schierhout G, Roberts I. Anti-epileptic drugs for preventing seizures following acute traumatic brain injury (Updated). *The Cochrane Library* Issue 4; 2001.
34. Shakespeare DT, Boggild M, Young C. Anti-spasticity agents for multiple sclerosis (Updated). *The Cochrane Library* Issue 4; 2001.
35. Kenyon S, Boulvain M, Neilson J. Antibiotics for preterm premature rupture of membranes (Updated). *The Cochrane Library* Issue 4; 2001.
36. Gulmezoglu AM, Hofmeyr GJ. Betamimetics for suspected impaired fetal growth (Updated). *The Cochrane Library* Issue 4; 2001.
37. van Vliet Hubertus AAM, Grimes DA, Helmerhorst Frans M, Schulz KF. Biphasic versus monophasic oral contraceptives for contraception (Updated). *The Cochrane Library* Issue 4; 2001.
38. Askie LM, Henderson-Smart DJ. Early versus late discontinuation of oxygen in preterm or low birth weight infants (Updated). *The Cochrane Library* Issue 4; 2001.
39. Osborn DA, Evans N. Early volume expansion for prevention of morbidity and mortality in very preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
40. Hofmeyr GJ. External cephalic version facilitation for breech presentation at term (Updated). *The Cochrane Library* Issue 4; 2001.
41. Davis PG, Henderson-Smart DJ. Extubation from low-rate intermittent positive airways pressure versus extubation after a trial of endotracheal continuous positive airways pressure in intubated preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
42. Singh M. Heated, humidified air for the common cold (Updated). *The Cochrane Library* Issue 4; 2001.
43. Hodnett ED. Home-like versus conventional institutional settings for birth (Updated). *The Cochrane Library* Issue 4; 2001.
44. Jones A, Rowe B, Peters J, Camargo C, Hammarquist C, Rowe B. Inhaled beta-agonists for asthma in mechanically ventilated patients (Updated). *The Cochrane Library* Issue 4; 2001.
45. Barrington KJ, Finer NN. Inhaled nitric oxide for respiratory failure in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
46. Davis PG, Henderson-Smart DJ. Intravenous dexamethasone for extubation of newborn infants (Updated). *The Cochrane Library* Issue 4; 2001.
47. Makrides M, Crowther CA. Magnesium supplementation in pregnancy (Updated). *The Cochrane Library* Issue 4; 2001.

48. Takken T, van der Net J, Helders PJM. Methotrexate for treating juvenile idiopathic arthritis (Updated). *The Cochrane Library* Issue 4; 2001.
49. Kirchmayer U, Davoli M, Verster A. Naltrexone maintenance treatment for opiod dependence (Updated). *The Cochrane Library* Issue 4; 2001.
50. Finer NN, Barrington KJ. Nitric oxide for respiratory failure in infants born at or near term (Updated). *The Cochrane Library* Issue 4; 2001.
51. Watson MC, Grimshaw JM, Bond CM, Mollison J, Ludbrook A. Oral versus intra-vaginal imidazole and triazole anti-fungal treatment of uncomplicated vulvovaginal candidiasis (thrush) (Updated). *The Cochrane Library* Issue 4; 2001.
52. Osborn DA. Thyroid hormone for preventing neurodevelopmental impairment in preterm infants (Updated). *The Cochrane Library* Issue 4; 2001.
53. Young GL, Jewell D. Topical treatment for vaginal candidiasis in pregnancy (Updated). *The Cochrane Library* Issue 4; 2001.

5

5

DEVELOPMENT OF AMSTAR: A MEASUREMENT TOOL TO ASSESS THE METHODOLOGICAL QUALITY OF SYSTEMATIC REVIEWS

Our objective was to develop an instrument to assess the methodological quality of systematic reviews by building upon previous tools, empirical evidence, and expert consensus.

A 37-item assessment tool was formed by combining 1) the enhanced overview quality assessment questionnaire (OQAQ), 2) a checklist created by Sacks, and 3) three additional items recently judged to be of methodological importance. This tool was applied to 99 paper-based and 52 electronic systematic reviews. Exploratory factor analysis was used to identify underlying components. The results were considered by methodological experts using a nominal group technique aimed at item reduction and design of an assessment tool with face and content validity.

The factor analysis identified 11 components. From each component, one item was selected by the nominal group.

A measurement tool to assess systematic reviews (AMSTAR) was developed. The AMSTAR tool consists of 11 items and has good face and content validity for measuring the methodological quality of systematic reviews. Additional studies are needed with a focus on the reproducibility and construct validity of AMSTAR before strong recommendations on its use can be made.

Beverley Shea, Jeremy M Grimshaw, George A Wells, Maarten Boers, Neil Andersson, Candyce Hamel, Ashley Porter, David Moher, Peter Tugwell, Lex M Bouter. Development of AMSTAR: A measurement tool to assess methodological quality of systematic reviews. BMC Medical Research Methodology 2007,7:10.

Background

It has been estimated that healthcare professionals attempting to keep abreast of their field would need to read an average of 17-20 original articles every day.[1] Increasingly, systematic reviews are being advocated as a way to keep up with current medical literature.[2] A well conducted systematic review addresses a carefully formulated question by analyzing all available evidence. It employs an objective search of the literature applying predetermined inclusion and exclusion criteria to the literature. As well, it critically appraises what is found to be relevant and extracts and synthesizes data from the available evidence base to formulate findings.[3]

However, in spite of the care with which they are conducted, systematic reviews may differ in quality and thus, yield different answers to the same question.[4] As a result, users of systematic reviews should be critical and look carefully at the methodological quality of the available reviews.[5]

A decade has passed since there has been any new development in tools to assess the quality of systematic reviews such as those created by Oxman and Guyatt[6] and Sacks.[7] There are now more than 26 instruments available to assess the quality of systematic reviews.[8] However, the majority of the available instruments are not widely used as they are lengthy and complicated to use. Furthermore, since their development, considerable empirical research has accumulated about potential sources of bias in systematic reviews. For example, recent methodological research has highlighted the potential importance of publication language and publication bias in systematic reviews.[9-11]

Therefore, our goal was to develop a new instrument for assessing the methodological quality of systematic reviews by building upon empirical data collected with previously developed tools and utilizing expert opinion.

This goal was pursued by two study objectives. Our first objective was to assess a large sample of systematic reviews using an item pool drawn from two available instruments used to assess methodological quality, supplemented by additional items judged to be needed on the basis of recent publication. We used exploratory factor analysis to identify the underlying component structure. Our second objective was to build on the results of this factor analysis by using experts in a nominal group technique (NGT) to reduce the items pool and to decide on a new assessment tool with face and content validity.

Methods

We designed a 37-item assessment tool that was developed by combining items from two available instruments: the enhanced overview quality assessment questionnaire (OQAQ)[8] containing 10 items and a checklist created by Sacks[7] containing 24 items. We supplemented this with three additional items based upon methodological advances in the field since the development of the original two instruments:
1) Language restriction: Language restriction in systematic reviews remains controversial. Some studies have suggested that systematic reviews that include only English language publications tend to overestimate effect sizes[10], whereas other studies suggest that such language restrictions may not do so.[11] An item was added to determine whether a language restriction was applied in selecting studies for the systematic review.
2) Publication bias: Publication bias refers to the tendency of research with negative findings being published less frequently, less prominently, or more slowly. It also refers to the tendency of research with positive findings being published more than once. Publication bias has been identified as a major threat to the validity of systematic reviews. Empirical research suggests that publication bias is widespread, though, a variety of methods are now available to assess publication bias.[12-19] We added an item to determine whether the authors assessed the likelihood of publication bias.

3) Publication status of studies suggests that published trials are generally larger and may show an overall greater treatment effect than those published in the 'grey' literature.[20] The importance of including the grey literature studies in systematic reviews has been addressed.[21] The assessment of the inclusion of grey literature considers whether or not the authors reported searching for grey literature. An item was added to ascertain whether the authors reported searching for grey literature.

Objective 1

The 37-item assessment tool was used to appraise 99 paper-based reviews identified from a database of reviews and meta-analyses[22] and 52 Cochrane systematic reviews from the Cochrane Database of Systematic Reviews.[9] Subsequent to the list of selected systematic reviews being generated, full copies of them were retrieved, copied, and masked to conceal author, institution, and journal. Reviews in languages other than English (i.e., French, German, and Portuguese) were translated into English with the assistance of colleagues before masking.[23]

For each included systematic review, two reviewers independently assessed the methodological quality with the 37 items (CH, BS).

Statistical analyses and graphs displaying the results obtained were produced using SPSS version 13.0 for Windows. The 37 items were subjected to principal components analysis and Varimax rotations were used to rotate the components. Items with low factor loadings of < 0.50 were removed.

Objective 2

We convened an international panel of 11 experts in the fields of methodological quality assessment and systematic reviews. The group was selected from three organizations involved both in the conduct of systematic reviews and in the assessment of methodological quality. The group was made up of clinicians, methodologists, epidemiologists, and new reviewers to the field. Some of these individuals were previously involved in the Cochrane Collaboration while a number were not. In examining the results of the factor analysis, they reflected critically on the components identified and decided on the items to be included in the new instrument. The nominal group process took place in San Francisco during a one day session.

We conducted the following NGT in order to achieve agreement. After delivery of an overview of the project and the planned process for the day, the panel reviewed the results of the factor analysis. The aim of the NGT was to structure interaction within the group. Firstly, each participant was asked to record his or her idea independently and privately. The ideas were then listed in a round-robin format; that is one idea was collected from each individual in turn and listed in front of the group by the facilitator. This process was continued until all ideas were listed. Individuals then privately recorded their judgements. Subsequent discussions then took place. The individual judgements were aggregated statistically to derive the group judgements. The nominal group was also asked to agree on a final label for each of the 11 components. A description was formulated for each of the items and a next-to-final instrument was assembled. This was then circulated electronically to the group for a final round of fine tuning.

Results

Objective 1

The items were subjected to factor analysis and only those items that loaded highly on one component (>.50) were retained. The described factor analysis made it possible to reduce the 37-item instrument to a shorter (29-items) instrument that measured 11 components (Table 1).

Objective 2

The nominal group discussed all 11 components (Table 1). The items most appropriate for the components as outlined in Instrument 3 were included in the draft instrument. The instrument is an 11-item questionnaire that asks reviewers to answer yes, no, can't answer, or not applicable. A separate question on language was identified in the factor analysis as a significant issue, but the nominal group felt that the contradictory evidence in the literature warranted removing this item from the shortened item list and capturing it under the question on publication status. Sources of support was one of the original components from the factor analyses. The group felt that conflict of interest was a more appropriate term for this component.

Discussion

Strengths and Weaknesses

Our purpose was to help users of systematic reviews to critically appraise systematic reviews. Therefore, we set out with a goal to develop a new instrument for assessing the methodological quality of systematic reviews by building upon empirical data on previously developed tools, empirical evidence, and utilizing expert opinion.

Since we had already created a dataset of 151 systematic reviews assessed using 37 completed items for each review, we were able to conduct a factor analysis as the first step in the creation of the new tool. A more commonly used approach would have been to harvest appropriate items from existing questionnaires. This method has been used extensively in the development of instruments for assessing the quality of both randomized and non-randomized studies of health care interventions.[24-26] The disadvantage of harvesting appropriate items from existing questionnaires is that it relies heavily on the validation of the source questionnaires.[27] Conducting a factor analysis made it possible to determine whether the measured dimensions could in principle be assessed using a smaller number of items.

Traditionally, factor analysis is divided into two types of analyses: exploratory and confirmatory. As its name indicates, exploratory factor analysis aims to discover the main constructs or dimensions of a concept by conducting a preliminary investigation of the correlations between all the identified variables. This process is also known as Principal Components Analysis (PCA). PCA has been recommended for use in test construction by Kline as a means of condensing the correlation matrix rather than as an aid to the interpretation of the factor-structure of a questionnaire.[28] Items with low factor loadings tend to be weakly correlated with other items and therefore, were removed. Various rotational strategies have also been proposed. The goal of each of them is to obtain a clear pattern of loadings, that is; factors that are somehow clearly marked by high loadings for some variables and low loadings for others.[29-30] We used this approach because it is useful when a body of theory or principles has been established but has not yet been operationalized into an evaluative framework.[31]

The structured-discussion format employed in this project enabled all participants to contribute to the refining of the assessment tool. The nominal technique followed involved experts, discussion, and a

consensus that was qualitative in nature. Consequently, it complemented the quantitative nature of factor analysis and as a result, the final tool had face and content validity as judged by the nominal consensus panel.

We recognize the need for further testing of AMSTAR. Additional studies are necessary with a focus on the reproducibility and construct validity of AMSTAR before strong recommendations on its use can be made. The Canadian Agency for Drugs and Technologies in Health (CADTH)[32] undertook an independent assessment of available quality assessment criteria for systematic reviews. Feedback from CADTH reviewers has been very positive. Further preliminary experience suggests that AMSTAR has good reliability and convergent validity which indicates that appraisers can apply it in a consistent way.

AMSTAR, if used widely after external validation, could also enable methodological research (i.e. meta-regression of item of AMSTAR and effect size of reviews). AMSTAR will remain a living document and advances in empirical methodological research will continue to improve the instrument. This is indeed likely to be the case with techniques to identify and quantify publication bias.[33] Although a number of alternative tests for publication bias exist, none have yet been validated.[34]

Publication bias remains an area of contention amongst those who assess the quality of systematic reviews. It remains a research priority. It is still not clear what the impact of publication bias is on making decisions in health care. We are aware of the 20 years of work dedicated to this area of research which has provided a few answers as to the effect publication bias may have on the overall results of estimating the impact of interventions.

Our instrument is an attempt to achieve consensus amongst current mainstream opinions. Inevitably, new evidence will modify current thinking in some areas and at that point, the AMSTAR will be updated.

Conclusions

A measurement tool to assess systematic reviews (AMSTAR) was developed. The tool consists of 11 items (Instrument 3) and has good face and content validity for measuring the methodological quality of systematic reviews. Additional studies are needed with a focus on the reproducibility and construct validity of AMSTAR before strong recommendations on its use can be made.

Acknowledgements

The authors would like to thank the following members of the San Francisco Nominal Group for their participation: Gail Kennedy, George Rutherford, George Wells, Peter Tugwell, Maarten Boers, Madhukar Pai, Kirby Lee, Maria Suarez-Almazor, Kaveh Shojania, Tara Horvath, and Kathryn McDonald. Thank you to Vivian Robinson and Adrienne Stevens for their helpful comments. The authors also thank Drs. Andy Oxman, Gord Guyatt and Henry Sacks for their methodological instruments and for providing additional information.

Table 1
Results from the factor analysis (37 items reduced to 29, and 11 components)

	Original instrument (item no)	1	2	3	4	5	6	7	8	9	10	11
1. Protocol	Sacks	.58										
2. Literature Search	Sacks				.82							
3. List of Trials Analyzed	Sacks		.75									
4. Log of Rejected Trials	Sacks								.68			
5. Treatment Assignment	Sacks		.80									
6. Ranges of Patients	Sacks											
7. Range of Treatment	Sacks		.88									
8. Range of Diagnosis	Sacks		.80									
9. Combinability Criteria	Sacks									.88		
10. Measurement	Sacks				.57							
11. Selection Bias	Sacks	.85										
12. Data abstraction	Sacks	.50										
13. Inter-observer Agreement	Sacks	.65										
14. Sources of Support	Sacks								.64			
15. Statistical Methods	Sacks			.81								
16. Statistical Errors	Sacks											
17. Confidence Intervals	Sacks			.73								
18. Subgroup Analysis	Sacks											
19. Quality Assessment	Sacks					.77						
20. Varying Methods	Sacks					.63						
21. Publication Bias	Sacks						.77					
22. Caveats	Sacks											
23. Economic Impact	Sacks											.84
24. Language I	Added to Sacks							.79				
25. Search Strategy	OQAQ (1)				.81							
26. Was the search comprehensive	OQAQ (2)											
27. Criteria used for deciding which studies to include	OQAQ (3)											
28. Was bias in the selection avoided	OQAQ (4)	.81										

Table 1
continued

29. Were the criteria used for assessing the validity reported	OQAQ (5)					.75						
30. Was the validity of all studies referred to in the text assessed using appropriate criteria	OQAQ (6)				.53							
31. Were the methods used to combine the finding of the relevant studies reported	OQAQ (7)											
32. Were the findings of the relevant studies combined appropriately	OQAQ (8)			.78								
33. Were the conclusions made by the author supported by the data	OQAQ (9)			.68								
34. Overall Summary	OQAQ (10)											
35. Publication Bias	Additional (1)						.80					
36. Publication Status	Additional (2)							.77				
37. Language II	Additional (3)							.63				

References

1. Davidoff F, Haynes B, Sackett D, Smith R. Evidence-based medicine: a new journal to help doctors identify the information they need. *BMJ* 1995; 310: 1085-86.
2. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998; 351: 123-27.
3. Systematic Review definition: http://www.nlm.nih.gov/nichsr/hta101/ta101014.html).
4. Moher D, Soeken K, Sampson M, Campbell K, Ben Perot L, Berman B. Assessing the quality of reports of systematic reviews in pediatric complementary and alternative medicine. *BMC Pediatr* 2002; 2(2).
5. Jadad A, Moher M, Browman G, Booker L, Sigouin C, Fuentes M. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *BMJ* 2000; 320: 537-40.
6. Oxman AD. Checklists for review articles. *BMJ* 1994; 309(6995): 648-51.
7. Sacks H, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. NEJM 1987; 316(8): 450-55.
8. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In *Systematic Reviews in Health Care*: Meta-analysis in context. Egger M, Smith GD, Altman DG, Eds. London: BMJ books 2001: 122-39.
9. *The Cochrane Library*. Chichester, UK: John Wiley & Sons Ltd, 2004; 3.
10. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997; 350: 326-29.
11. Moher D, Pham B, Klassen T, Schulz K, Berlin J, Jadad A, Liberati A. What contributions do languages other than English make to the results of meta-analyses? *J Clin Epidemiol* 2000; 53: 964-72.
12. Pai M, McCulloch M, Colford J Jr., Bero LA. Assessment of Publication Bias in Systematic Reviews on HIV/AIDS. (http://www.igh.org/Cochrane/pdfs/MSRI_workshop_talk_abstract.pdf)
13. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; 263: 1385-89.
14. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997; 9: 15-21.
15. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence the estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000; 356:1228-31.
16. Pham B, Platt R, McAuley L, Sampson M, Klassen T, Moher D. Detecting and minimizing publication bias. A systematic review of methods. Technical Report, Thomas C. Chalmers Centre for Systematic Reviews, Ottawa, Canada 2000.
17. Stern JM, Simes R. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997; 315: 640-45.
18. Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis- power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000; 53:1119-29.
19. Sutton A, Duval S, Tweedie R, Abrams K, Jones D. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000; 320: 1574-77.
20. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. In *The Cochrane Library*, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.
21. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence the estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000; 356: 1228-31.
22. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of randomized controlled trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999; 3(12): i-iv, 1-98.
23. Shea B. Assessing the reporting quality meta-analyses of randomized controlled trials. MSc thesis. University of Ottawa, Department of Epidemiology and Community Medicine;1999.

24. Downs S, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998; 52: 377-84.
25. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; 17: 1-12.
26. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The Delphi list: a consensus list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998; 51: 1235-41.
27. Cluzeau FA. Development and application of an appraisal instrument for assessing the methodological quality of clinical practice guidelines in the United Kingdom. London, England: St.George's Hospital Medical School, University of London; 2001.
28. Kline P. An Easy Guide to factor Analysis. London: Routledge; 1994.
29. Norman GR, Streiner DL. Biostatistics: The Bare Essentials, 2nd edition. St. Louis: Mosby; 2000.
30. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use, 3rd edition. UK. Oxford University Press; 2003.
31. McDowell I, Jenkinson C. Development standards for health measures. *Journal of Health Services Research and Policy* 1996; 1(4): 238-46.
32. Canadian Agency for Drugs and Technologies in Health (www.cadth.ca).
33. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ* 2006; 333: 597-600.
34. Rothstein HR, Sutton AJ, Borenstein M, eds. Publication bias in meta-analysis: prevention, assessment and adjustments. Sussex: John Wiley and Sons; 2005.

6

6

INTERNAL VALIDATION OF AMSTAR: A MEASUREMENT TOOL TO ASSESS SYSTEMATIC REVIEWS.

Our purpose was to measure the agreement, reliability, construct validity and feasibility of AMSTAR (A Measurement Tool to Assess Systematic Reviews).

We randomly selected 30 systematic reviews from a data-base. Each was assessed by two reviewers using: 1) the enhanced quality assessment questionnaire (OQAQ); 2) Sacks instrument; and 3) our newly developed measurement tool (AMSTAR). We report on reliability (inter-observer kappas of the 11 AMSTAR items), intra-class correlation coefficients (ICCs) of the sum scores, construct validity (ICCs of the sum scores of AMSTAR compared to that of other instruments), and completion times.

The inter-rater agreement of the individual items of AMSTAR was substantial with a mean kappa of 0.70 (95% CI: 0.57; 0.83) (range 0.38-1.0). Kappas recorded for the other instruments were 0.63 (0.38; 0.78) for enhanced OQAQ and 0.40 (0.29; 0.50) for the Sacks' instrument. The ICC of the total score for AMSTAR was 0.84 (0.65; 0.92); compared with 0.91 (0.82; 0.96) for OQAQ and 0.86 (0.71; 0.94) for the Sacks instrument. AMSTAR proved easy to apply, each review taking about 10-15 minutes to complete.

AMSTAR has good agreement, reliability, construct validity and feasibility. These findings need confirmation by a broader range of assessors and a more diverse range of reviews.

Beverley J Shea, Candyce Hamel, Lex M Bouter, Betsy Kristjansson, Jeremy M Grimshaw, David Henry, Maarten Boers. Internal validation of AMSTAR: a measurement tool to assess systematic reviews. (Journal of Clinical Epidemiology).

Background

Systematic reviews have become the standard methodology to assess and summarize applied health research, but the quality of the reviews themselves has received relatively little attention. Quality can be defined as the likelihood that the design of a systematic review will generate unbiased results.[1] Systematic reviews have appeared in medical journals since the late 1970s. Thousands of systematic reviews have become available on all areas of health care and a substantial portion of them have been produced by the Cochrane Collaboration. High quality in systematic reviews is needed to support a valid interpretation of review findings.

In a previous study we summarized the literature, tested available tools, and reached the conclusion that current instruments for conducting methodological quality assessments of systematic reviews were suboptimal and needed improvement.[2, Chapter 2] We then empirically developed a new measurement tool.[3, Chapter 5] Briefly, the two best available instruments to assess quality were applied to 151 systematic reviews of interventions. These reviews included both meta-analyses and qualitative reviews. Their scores on individual items were subjected to factor analysis to examine the extent to which the items measure one or more common themes. Factor analysis thus facilitated the elimination of redundant items. On this basis, we designed draft items for the new instrument. This draft was the topic of a nominal group consensus technique involving experts. The resulting instrument was subjected to a small pilot test that lead to minor refinements, resulting in the final version of AMSTAR (Instrument 1). This new instrument is an 11-item questionnaire requiring assessors to answer yes, no, can't answer or not applicable.[3, Chapter5]

The present study concerns the internal validation of a measurement tool to assess systematic reviews (AMSTAR) using the set of reviews employed in its development. We will focus on parameters of agreement, reliability, construct validity and feasibility through comparisons with the other instruments.

Methods

We used a computer generated random sample of 30 of the previously reviewed 151 systematic reviews.[2, Chapter 5] This sample contained 11 Cochrane and 19 non-Cochrane reviews, including meta-analyses and qualitative reviews. The topics of the reviews ranged across the spectrum of medicine.[4-33] Two reviewers (one without formal training) applied the new AMSTAR instrument and the two quality assessment tools, the enhanced overview quality assessment questionnaire (OQAQ), and the Sacks instrument to all 30 reviews (CH, BS). For each reviewer, the dataset extracted contained three quality ratings for each review, yielding a total of six ratings per review.

Agreement and Reliability

We calculated overall agreement and Cohen's kappa for each item ('yes' scores vs. any other scores).[34] Bland and Altman's limits of agreement methods explained agreement graphically.[35-37] We awarded each item scoring 'yes' one point and added these to calculate a total score. Intra-class correlation coefficients assessed reliability of this total score.[38] We further scrutinized items and reviews with kappas below 0.50. Finally, using new assessments we repeated the exercise for the OQAQ and Sacks instruments. Kappa values of less than 0 rate as less than chance agreement; 0.01-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; and 0.81-0.99 almost perfect agreement.[34]

Construct validity

The new instrument already has high face and content by virtue of its construction process.[3] In the current study, we assessed construct validity by converting the mean total score (mean of two raters CH and BS) of each of the 30 reviews to a percentage of the maximum score for each of the three instruments. Intra-class correlation coefficients then assessed convergence of the total scores between each pair of instruments (AMSTAR-OQAQ, AMSTAR-Sacks, and OQAQ-Sacks).

Feasibility

Based on a guideline for assessing feasibility of instrument use developed by the OMERACT group (Outcome Measures in Rheumatology),[39] we compared the feasibility of the new instrument to that of the existing instruments by recording the time it took to complete scoring and the instances where scoring was difficult or impossible. The wording of individual items is critical for the performance of AMSTAR and fine tuning is expected to be an ongoing task.

SPSS (version 13) and MedCalc were used to analyze the data and the results were expressed as means and 95% confidence intervals unless otherwise noted.

Results

The sample of 30 reviews adequately covered a wide range of quality, albeit with some under-representation of poor quality reviews. Overall quality scores on AMSTAR ranged from 3 to 10 (out of a maximum of 11) with a flat distribution between 3.5 and 10 and a mean percentage score of 49.4%. The overall quality scores on Sacks ranged from 5 to 16 (out of a maximum score of 24) a mean percentage score of 41.6% and for OQAQ scores ranged from 3 to 10 (out of a maximum score of 10) with a mean percentage score of 63.3%.

Agreement and Reliability

The inter-observer agreement of the individual items in the AMSTAR was high: mean 0.88 (range 0.73-1.0) with a mean kappa of 0.70 (0.57; 0.83) (range 0.38-1.0). However, items 4 (publication status), 7 (report of assessment of scientific quality) and 9 (appropriate method to combine studies) scored fair to moderate at 0.38, 0.42, and 0.45, respectively. On the first two of these items, overall agreement was substantial at 0.80 and the relatively low kappa may be explained by a skewed distribution, i.e. a high number of reviews in which the reviewers agreed on the score 'no' (item 4) and 'yes' (item 7), respectively. On item 8, overall agreement was also satisfactory at 0.74. Compared to the other instruments, agreement on individual items was similar to OQAQ: mean kappa 0.63 (0.39; 0.78) (range 0.39-0.84), and superior to the Sacks instrument: mean kappa 0.40 (0.29; 0.50) (range 0.47- +0.93). In these instruments, fair to moderate agreement was also seen in the items covering assessment of scientific validity, statistical combinability and comprehensively literature searching (Table 1).

For the AMSTAR total score the mean difference between the two observers' scores was 0.2 (0.36 to 0.91). Agreement was similar in reviews with high and low quality scores (Figure 1).

The inter-observer ICC for the total score was excellent for all instruments: AMSTAR 0.84 (95% CI: 0.65 to 0.92); OQAQ 0.91 (95% CI: 0.82 to 0.96) Sacks instrument 0.86 (95% CI: 0.71 to 0.94). In one non-Cochrane review 12 observers differed by three points (6 vs. 9). In this review, the differences were noted on AMSTAR questions addressing duplication study selection and data extraction (item 2), publication status (item 4) and methods used to combine studies (item 9). In one Cochrane review, 15 observers differed by four points (1 vs. 5). In this review, differences were noted on AMSTAR

questions assessing the a priori design (item 1), publication status (item 4), scientific quality (item 7) and methods used to combine studies (item 9). The overall quality of Cochrane reviews included in this dataset was somewhat higher than non-Cochrane reviews.

The qualitative analysis of the data on agreement led us to make minor modifications to the wording of some items. In particular, under the item publication bias, the wording was changed to clarify the purpose of the question, that is: to ask whether or not the status of publication was used as an inclusion criterion (see item 4, instrument 3). Additional available electronic databases were also added to the question on literature searching and minor adjustments were made to the wording for the item on methods used to combine findings.

Construct validity

Expressed as a percentage of the maximum score, the results of AMSTAR showed convergence with the results of the other instruments. ICCs were 0.66 (95% CI: 0.28 to 0.84) against OQAQ and 0.83 (95% CI: 0.64 to 0.92) against Sacks' instrument. The ICC obtained when comparing OQAQ to Sacks was 0.86 (95% CI: 0.70 to 0.93).

Feasibility

AMSTAR proved easy to apply, each review taking about 10-15 minutes to complete. OQAQ took on average more than 20 minutes to complete, and Sacks over 40 minutes. At least two of the reviewers expressed difficulty with scoring item 4 on publication status: 'was the status of publication (i.e. grey literature) used as an inclusion criterion?'

Discussion

This study suggests that AMSTAR has good agreement, reliability, construct validity and feasibility to assess the quality of systematic reviews. AMSTAR was compared to the two best currently available tools. Its performance in terms of agreement and reliability was similar to OQAQ and better than Sacks; it adds relevant items not present in either instrument (e.g. publication status, conflict of interest) and has better feasibility than OQAQ or Sacks. We think AMSTAR can be applied to a wide variety of systematic reviews but recognize that it has only been tested on systematic reviews of randomized control trials evaluating treatment interventions.

Careful consideration was given to the wording of the individual items and minor adjustments were made. Despite this process, agreement between observers was disappointingly low on three items. One of these items assessed publication restriction. After discussion between observers, we reworded the descriptor slightly and feel this will improve agreement. The other two items describe 'report of assessment of scientific quality' and 'appropriate method to combine studies.' We feel these items would be difficult to score regardless of their wording. Accordingly, agreement was low on similar items in the other instruments as well. Subjective judgment comes into play when one is asked to assess whether quality of included studies was assessed adequately. Conceivably, one could increase reliability of assessment by providing more detailed instructions or by adding more items or criteria. This would however, decrease feasibility. It should also be noted that overall agreement on these items was good, so their relatively low kappa's are likely caused by skewness in the responses, i.e. a majority of responses in either the 'yes' or the 'no' category. This is a well-known limitation of the kappa statistic.

Although similar and excellent, we assume the reliability of total score for AMSTAR and OQAQ is due to the fact that the raters were familiar with both existing instruments and have been using them for

several years. Scrutiny of the two reviews with the most discrepant scores revealed that in both, the assessors at least agreed on the placement of the quality score above or below 50%. Finally, AMSTAR showed good (convergent) construct validity in comparison with the two existing instruments.

We feel the main advantage of AMSTAR above OQAQ and Sacks' instrument lies in its better compromise between comprehensiveness and feasibility. It adds relevant dimensions to those covered in the OQAQ without becoming unwieldy like the Sacks instrument. In addition, a recently published study concluded that the underlying construction of OQAQ is designed for the assessment of meta-analyses. Thus, it is almost impossible for any other type of review to score highly on the OQAQ.[40]

Feasibility of the AMSTAR is documented in terms of the time required to complete an assessment while using it: about 10-15 minutes, which is approximately half the time (or less) needed to complete the other instruments.

Our study has some limitations. We did not compare AMSTAR to the current state-of-the-art reporting quality of meta-analysis (QUOROM). The reason for this is that QUOROM is not specifically designed to assess methodological quality. Rather, it is specifically focused on the reporting quality of the review. This does not detract from the utility of QUOROM but its limited focus made it unsuitable for our study.

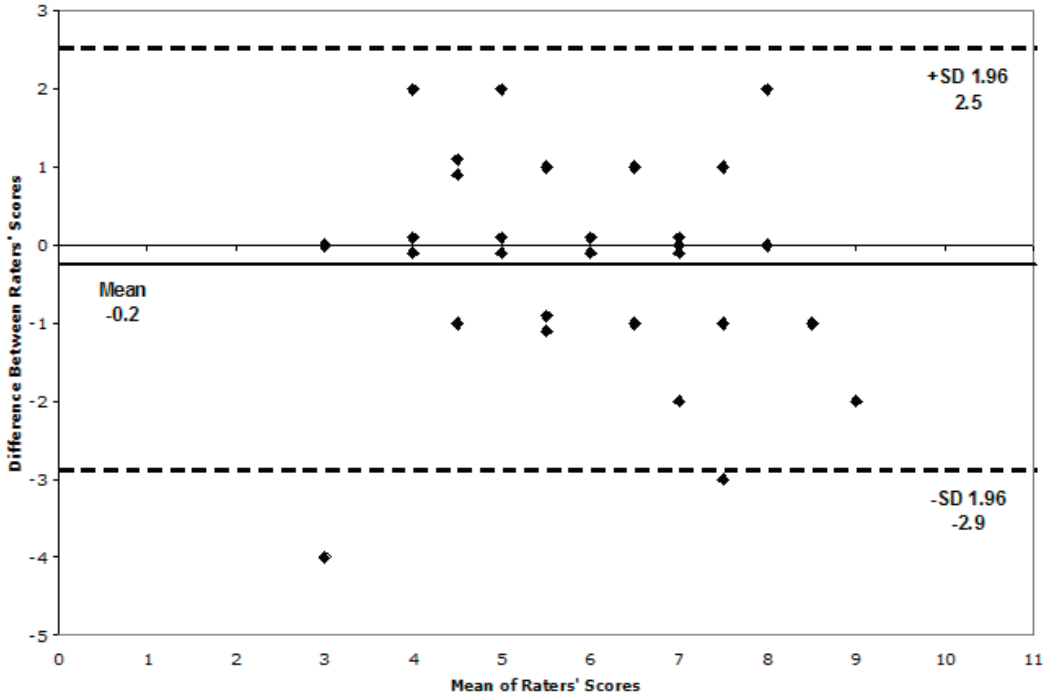
A second limitation of the present study is the fact that the sample of reviews used is derived from the source used to develop AMSTAR and that one of the assessors is the principal investigator. Thus, application to other reviews and by other assessors is necessary to discover the full potential of this tool. Finally, the number of reviews used to validate AMSTAR was rather small.

In summary, AMSTAR is an empirically-developed instrument for documenting the quality of systematic reviews. It was found to have good agreement, reliability and construct validity in a limited test setting. It combines in one instrument a level of comprehensiveness and feasibility not found in existing instruments. We encourage others to test our new instrument on other samples of systematic reviews. Its ongoing application in the assessment of the quality of systematic reviews will provide further confirmation of its utility.

Table 1
Assessment of the inter-rater agreement for AMSTAR

Items	Kappa (95% CI)
1. Was an 'a priori' design provided?	0.80 (0.63 to 0.90)
2. Was there duplicate study selection and data extraction?	0.80 (0.17 to 0.81)
3. Was a comprehensive literature search performed?	0.72 (0.40 to 0.87)
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?	0.38 (0.28 to 0.70)
5. Was a list of studies (included and excluded) provided?	0.56 (0.07 to 0.79)
6. Were the characteristics of the included studies provided?	0.74 (0.45 to 0.86)
7. Was the scientific quality of the included studies assessed and documented?	0.42 (0.23 to 0.72)
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	0.74 (0.45 to 0.87)
9. Were the methods used to combine the findings of studies appropriate?	0.45 (0.12 to 0.70)
10. Was the likelihood of publication bias assessed?	0.88 (0.75 to 0.94)
11. Were potential conflicts of interest included?	0.92 (0.83 tp 0.96)

Figure 1
Bland and Altman plot of inter-rater agreement on AMSTAR total score



References

1. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995; 16(1): 62-73.
2. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In: Egger M, Smith GD, Altman DG, Eds. Systematic Reviews in Health Care: Meta-analysis in context. London: *BMJ Books* 2001; 122-39. (Chapter 2)
3. Shea B, Grimshaw J, Wells G, Boers M, Andersson N, Hamel C, Porter A, Moher D, Tugwell P, Bouter L. Development of AMSTAR: A Measurement Tool to Assess Reviews Methodological Quality of Systematic Reviews. *BMC Medicial Research Methodology* 2007; 7: 10. (Chapter 3)
4. Anonymous. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. Early Breast Cancer Trialists' Collaborative Group. *NEJM* 1989; 319(26): 1681-92.
5. Appel LJ, Miller ER, Seidler AJ, Whelton PK. Does Supplementation of Diet with 'Fish Oil' Reduce Blood Pressure. *Arch Intern Med* 1993; 153: 1429-38.
6. Buring JE, Evans DA, Mayrent SL, Rosner B, Colton T, Hennekens CH. Randomized trials of aminoglycoside antibiotics: Quantitative overview. *Rev Inf Dis* 1988; 10(5): 951-57.
7. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *NEJM* 1977; 297(20): 1091-96.
8. Clagett GP, Reisch JS. Prevention of venous thromboembolsm in general surgical patients. Results of meta-analysis. *Annals of Surgery* 1988; 208(2): 227-40.
9. Counsell C, Warlow C, Naylor R. Different patches in carotoid surgery. *The Cochrane Library* 1996; Issue 3.
10. Daya S. Comparison of FSH and HMG in IVF. *The Cochrane Library* 1996; Issue 3.
11. Duley L, Gulmezoglu AM, Henderson-Smart DJ. Anticonvulsants for pre-eclampsia. *The Cochrane Library* 1996; Issue 3.
12. Fanning J, Bennett TZ, Hilgers RD. Meta-analysis of Cisplatin, Doxorubicin, and Cyclophosphamide Versus Cisplatin and Cyclophosphamide Chemotherapy of Ovarian Carcinoma. *Obstet Gynecol* 1992 Dec; 80(6): 954-60.
13. Gent M, Roberts RS. A meta-analysis of the studies of dihydroergotamine plus heparin in the prophylaxis of deep vein thrombosis. *Chest* 1986; 89(5): 396S-400S.
14. Gotzsche PC, Gjorup I, Bonnen H, Brahe NE, Becker U, Burcharth F. Somatostatin vs placebo in bleeding oesophageal varices: Randomised trial and meta-analysis. *BMJ* 1995; 310(6993): 1495-98.
15. Graves P. Malaria vaccines. *The Cochrane Library* 1996; Issue 3.
16. Henderson WG, Goldman S, Copeland J, Moritz TE, Harker LA. Antiplatelet or anticoagulant therapy after coronary artery bypass surgery: A meta-analysis of clinical trials. *Ann Intern Med* 1989; 111(9): 743-50.
17. Hodnett ED. Alternative versus conventional delivery settings. *The Cochrane Library* 1996; Issue 3.
18. Hofmeyr GJ. Abdominal decompression. *The Cochrane Library* 1996; Issue 3.
19. Hopfenmuller W. Nackweis der therapeutischen Wirksamkeit eines Ginkgo biloba-Spezialextrakes: Meta-Analyse von 11 klinischen Studien bei Patienten mit Hirnleistungsstorungen im Alter. *Arzneim-Forsch* 1994; 44(9): 1005-13.
20. Hughes E, Fedorkow DM, Daya S, Sagle MA, van de Kopple P, Collins JA. The routine use of gonadotropin-releasing hormone agonists prior to in vitro fertilizatoin and gamete intrafollopian transfer: A meta-analysis of randomized controlled trials. *Fertil Steril* 1992; 58(5): 888-96.
21. Kaufmann PG, Jacob RG, Ewart CK, Chesney MA, Muenz LR, Doub N, Mercer W. Hypertension Intervention Pooling Project. *Health Psychol* 1988; 7 Suppl: 209-24.
22. Kramer MS. Maternal antigen avoidance as lactation. *The Cochrane Library* 1996; Issue 3.

23. Lycka BA. Postherpetic neuralgia and systemic corticosteroid therapy. Efficacy and safety. *Int J Dermatol* 1990; 29(7): 523-27.
24. McGrath JJ, Soares KVS. Tardive dyskinesia and benzodiazepines. *The Cochrane Library* 1996; Issue 3.
25. Mulrow CD, Mulrow JP, Linn WD, Aguilar C, Ramirez C. Relative efficacy of vasodilator therapy in chronic congestive heart failure. Implications of randomized trials. *JAMA* 1988; 259(23): 3422-26.
26. Ohlsson A. Treatments of preterm premature rupture of the membranes: a meta-analysis. *Am J Obstet Gynecol* 1989 Apr; 160(4): 890-906.
27. Perez-Escamilla R, Pollitt E, Lonnerdal B, Dewey KG. Infant Feeding Policies in Maternity Wards and Their Effect on Breast-Feeding Success: An Analytical Overview. *Am J Public Health* 1994; 84: 89-97.
28. Renfrew MJ, Lang S. Breastfeeding and discharge times. *The Cochrane Library* 1996; Issue 3.
29. Renfrew MJ, Lang S. Breastfeeding and early contact. *The Cochrane Library* 1996; Issue 3.
30. Soares KVS, McGrath JJ, Deeks JJ. Tardive dyskinesia and GABA agonist drugs. *The Cochrane Library* 1996; Issue 3.
31. Thacker SB. Quality of controlled clinical trials. The case of imaging ultrasound in obstetrics: a review. *BJOG* 1985; 92(5): 437-44.
32. Velanovich V. Crystalloid versus colloid fluid resuscitation: a meta-analysis of mortality. *Surgery* 1989; 105(1): 65-71.
33. Wilson APR., Shrimpton S, Jaderberg M. A meta-analysis of the use of amoxycillin-clavulanic acid in surgical prophylaxis. *J Hosp Infect* 1992 Nov; 22(Suppl A): 9-21.
34. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960; 0(1): 622-26.
35. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement 1986; *Lancet* i: 307-10. http://www-users.york.ac.uk/~mb55/meas/ba.htm
36. Bland JM, Altman DG. Statistical methods for assessing agreement between measurements. *Biochimica Clinica* 1987; 11: 399-404.
37. Bland JM, Altman DG. This week's citation classic: Comparing methods of clinical measurement. *Current Contents* 1992; CM20(40): 8.
38. Uebersax JS. Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin* 1987; 101: 140-46.
39. Tugwell P, Bombardier C. A methodological framework for developing and selecting endpoints in clinical trials. *J Rheumatol* 1982; 9(5): 758-62.
40. Biondi-Zoccai G, Lotrionte M, Abbate A, Testa L. Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. *BMJ* 2006; 332: 202-6.

7

7

EXTERNAL VALIDATION OF A MEASUREMENT TOOL TO ASSESS SYSTEMATIC REVIEWS (AMSTAR)

Thousands of systematic reviews have been conducted in all areas of health care. However, the methodological quality of these reviews is variable and the risk of bias should routinely be appraised. A measurement tool to assess systematic reviews (AMSTAR) is a new instrument to assess the methodological quality of systematic reviews.

AMSTAR was used to appraise 42 systematic reviews focusing on therapies to treat gastro-esophageal reflux disease, peptic ulcer disease, and other acid-related diseases. Two assessors applied the AMSTAR to each review. Two other assessors, plus a clinician and/or methodologist applied a global assessment to each review independently.

The sample of 42 reviews covered a wide range of methodological quality. The overall scores on AMSTAR ranged from 0 to 10 (out of a maximum of 11) with a mean of 4.6 (95% CI: 3.7 to 5.6) and median 4.0 (range 2.0 to 6.0). The inter-observer agreement of the individual items ranged from moderate to almost perfect agreement. Nine items scored a kappa of > 0.75 (95% CI: 0.55 to 0.96) (Φ> 0.76). Item 4 had a kappa of 0.64 (95% CI: 0.40 to 0.88) (Φ =0.64) and item 8 a kappa of 0.51 (95% CI: 0.25 to 0.78) (Φ =0.56). The reliability of the total AMSTAR score was excellent (kappa 0.84 (95% CI: 0.67 to 1.00) (Φ=0.85) and Pearson's R 0.96 (95% CI: 0.92 to 0.98).

The overall scores for the global assessment instrument ranged from 2 to 7 (out of a maximum score of seven) with a mean of 4.43 (95% CI: 3.6 to 5.3) and median 4.0 (range 2.25 to 5.75). The agreement for the global assessment was lower with a kappa of 0.63 (95% CI: 0.40 to 0.88). Construct validity was shown by AMSTAR convergence with the results of the global assessment instrument: Pearson's Rank Correlation Coefficient 0.72 (95% CI: 0.53 to 0.84). For the AMSTAR total score, the limits of agreement were -0.19 ±1.38. This translates to a minimum detectable difference between reviews of 0.64 'AMSTAR points'. The AMSTAR instrument has good inter-observer reliability and validity. Further validation efforts should also concern perceived utility by review authors and end users of reviews.

B Shea, LM Bouter, J Peterson, M Boers, Z Ortiz, N Andersson, T Ramsay, A Bai, V Shukla, JM Grimshaw. External validation of a measurement tool to assess systematic reviews (AMSTAR) PLoS ONE 2(12): e1350. doi:10.1371/journal.pone.0001350

Introduction

High quality systematic reviews are increasingly recognized as providing the best evidence to inform health care practice and policy.[1] The quality of a review, and so its worth, depends on the extent to which scientific review methods were used to minimize the risk of error and bias. The quality of published reviews can vary considerably, even when they try to answer the same question.[2] As a result, it is necessary to appraise their quality (as is done for any research study) before the results are implemented into clinical or public health practice. Much has been written on how best to appraise systematic reviews. While there is some variation on how this is achieved, most agree on key components of the critical appraisal.[3] Methodological quality can be defined as the extent to which the design of a systematic review will generate unbiased results.[4]

Several instruments exist to assess the methodological quality of systematic reviews[5], but not all of them have been developed systematically or empirically validated and have achieved general acceptance. The authors of this paper acknowledge that the methodological quality and reporting quality for systematic reviews is very different. The first, methodological quality, considers how well the systematic review was conducted (literature searching, pooling of data, etc.). The second, reporting quality, considers how well systematic reviewers have reported their methodology and findings. Existing instruments often try to include both types of methods without being conceptually clear about the differences.

In an attempt to achieve some consistency in the evaluation of systematic reviews we have developed a tool to assess their methodological quality. This builds on previous work[6] and is based on empirical evidence and expert consensus. A measurement tool to assess systematic reviews (AMSTAR) was highly rated in a recent review (personal communication) of quality assessment instruments performed by the Canadian Agency for Drugs and Technologies in Health (CADTH). In this study we present the results of an external validation of AMSTAR using data from a series of systematic reviews obtained from the gastroenterology literature.

Methods

The characteristics and basic properties of the instrument have been described elsewhere.[7] Briefly, a 37-item initial assessment tool was formed by combining a) the enhanced overview quality assessment questionnaire (OQAQ) scale, b) a checklist created by Sacks, and c) three additional items recently judged by experts in the field to be of methodological importance. In its development phase the instrument was applied to 99 paper-based and 52 electronic systematic reviews.[6, 7] Exploratory factor analysis was used to identify underlying components. The results were considered by methodological experts using a nominal group process to reduce the number of items and design an assessment tool with face and content validity. This process lead to an 11-item instrument.[7] A description of the instrument is provided in instrument 1.

External validity

For our validation test set we chose to use systematic reviews or meta-analyses in the area of gastroenterology, specifically upper gastrointestinal. CADTH's informational specialist searched electronic bibliographic databases (i.e. Medline, Central and EMBASE) up to and including 2005. A total of 42 systematic reviews met the a priori criteria and were included.[8] This sample included seven electronic Cochrane systematic reviews and 35 paper-based non-Cochrane reviews. The topics of the reviews ranged across the spectrum of GI problems such as dyspepsia, gastro-esophageal reflux disease (GERD), peptic ulcer disease (PUD), and also GI drug interventions such as H2 receptor antagonists and proton pump inhibitors.[9-50]

Two CADTH assessors independently applied AMSTAR to each review and reached agreement on the assessment results. To assess construct validity, two reviewers (JP, ZO) plus a clinician and/or methodologist (MB, DF, DP, MO, and DH) applied a global assessment to each review.[51] (Annex 1)

Agreement and reliability

We calculated an overall agreement score using the weighted Cohen's kappa, as well as one for each item.[52] (Table 1) Bland and Altman's limits of agreement methods were used to display agreement graphically[53, 54] (Fig. 1). We calculated the percentage of the theoretical maximum score. Pearson's Rank correlation coefficients were used to assess reliability of this total score. For comparisons of rating the methodological quality we calculated chance-corrected agreement (using kappa) and chance-independent agreement (using Φ).[52, 55, 56] We accepted a correlation of >0.66. We further scrutinized items and reviews with kappa scores below 0.66.[52] Kappa values of less than 0 rate as less than chance agreement; 0.01-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; and 0.81-0.99 almost perfect agreement.[52, 57] We calculated PHI Φ for each question.[55, 58]

Construct validity

We assessed construct validity (i.e. evaluation of a hypothesis about the expected performance of an instrument) by converting the total mean score (mean of the two assessors) for each of the 42 reviews to a percentage of the maximum score for AMSTAR and of the maximum score of the global assessment instrument. We used Pearson's Rank correlation coefficients, Pearson's R and Kruskal-Wallis test to further explore the impact of the following items on the construct validity of AMSTAR: a) Cochrane systematic review vs. non-Cochrane systematic reviews[59, 60], b) journal type[61], c) year of publication[62], d) conflict of interest[63], e) impact factor[63], and f) number of pages.[64] We studied these in the context of a priori hypotheses concerning the correlation of AMSTAR scores. Because of the nature of their development, we anticipated that Cochrane systematic reviews would have higher quality scores than non-Cochrane systematic reviews and those electronic or general journals would score higher than specialist journals. We reported on impact factors for these journals. We hypothesized that reviews published more recently would be of higher quality than those published earlier. In addition, we anticipated that reviews declaring a conflict of interest might have lower quality scores.[63, 64]

We assessed the practicability of the new instrument by recording the time it took to complete scoring and the instances where scoring was difficult. We interviewed assessors (N=6) to obtain data on clarity, ambiguity, completeness and user-friendliness.

We used SPSS (versions 13 and 15) and MedCalc for Windows, version 8.1.0.0.

Results

The 42 reviews included in the study had a wide range of quality scores. The overall scores estimated by the AMSTAR instrument ranged from 0 to 10 (out of a maximum of 11) with a mean of 4.6 (95% CI: 3.7 to 5.6; median 4.0 (range 2.0 to 6.0). The overall scores for the global assessment instrument ranged from 2 to 7 (out of a maximum score of seven) with a mean of 4.43 (95% CI: 3.6 to 5.3) and median 4.0 (range 2.5 to 5.3).

Agreement and Reliability

The reliability of the total AMSTAR score between two assessors (the sum of all items answered 'yes' scored as 1, all others as 0) was (kappa 0.84 (95% CI: 0.67 to 1.00, Φ=0.85) and Pearson's R 0.96 (95% CI: 0.92 to 0.98). The inter-rater agreement (kappa) between two raters for the global assessment was 0.63 (95% CI: 0.40 to 0.88).

Items in AMSTAR displayed levels of agreement that ranged from moderate to almost perfect; nine items scored a kappa of > 0.75 (0.55 to 0.96 (and Φ > 0.76). Item 4 had a kappa of 0.64 (0.40 to 0.88) Φ=0.64 and item 8 a kappa of 0.51(0.25 to 0.78 Φ=0.56). The reliability of the total AMSTAR score was excellent (kappa 0.84 (95% CI: 0.67 to 1.00 and Pearson's R 0.96 (95% CI: 0.92 to 0.98). For the AMSTAR total score, the limits of agreement were -0.19 ±1.38 (Fig. 1).

The mean age of our reviewers was 40.57, median 43. Fifty-seven percent were identified as experts in methodology and 43% were identified as content experts in the field.

Construct validity

Expressed as a percentage of the maximum score, the results of AMSTAR converged with the results of the global assessment instrument (Pearson's Rank Correlation Coefficient 0.72 (95% CI: 0.53 to 0.84)). AMSTAR scoring also upheld our other a priori hypotheses. The sub-analysis revealed that Cochrane reviews had significantly higher scores than paper-based reviews with a (R=37.21 n=7) for Cochrane reviews and (R=18.36 n=35) for paper-based (P < 0.0002). Cochrane reviews (R= 37.21 n=7) also scored higher than reviews published in general journals (R=25.77 n=11) and specialty journals (R=14.96, n=24) (P < 0.0001). Reviews published from 2000 onward had higher AMSTAR scores than earlier reviews (R=25.20, n=25 vs. R=13.12, n=17; P =0.0002).

The journals had the following overall summary statistics for the impact factors: mean 5.88 (95% CI: 3.9 to 7.9) median 3.3 (lowest value 1.4, highest value 23.9). There is no statistical association between AMSTAR score and impact factor (Pearson's R (0.555 P=0.7922)). There was however a significant association found with the number of pages and AMSTAR scores (Pearson's R (0.5623 P=0.0001 n=42). We found no association (R 0.1773 P=0.0308) when we removed the outliers (i.e. systematic reviews with higher page numbers).

Conflict of interest was poorly presented. Of the 42 reviews assessed, no study had appropriately declared their conflict of interest. Therefore, we were unable to assess whether or not funding had a positive or negative effect on the AMSTAR score.

Practicability

Both AMSTAR and the global assessment required on average 15 minutes to complete, but with the latter, assessors expressed difficulty in reaching a final decision in the absence of comprehensive guidelines. In contrast, AMSTAR was well received.

Discussion

Principal findings

This paper describes an external validation of AMSTAR. This new measurement tool to assess methodological quality of systematic reviews showed satisfactory inter-observer agreement, reliability and construct validity in this study. Items in AMSTAR displayed levels of agreement that ranged from moderate to almost perfect. The reliability of the total AMSTAR score was excellent. Construct validity was shown by AMSTAR convergence with the results of the global assessment instrument.

We found a significant association between number of published pages and overall AMSTAR score, suggesting that the longer the manuscript, the higher the quality score. It should be interpreted with caution given the fact that only a couple of the longer reviews largely drive the hypothesis tests. We found no association when the outliers were removed from the dataset. We did not find an association between AMSTAR score and impact factor.

The AMSTAR instrument was developed pragmatically using previously published tools and expert consensus. The original 37 items were reduced to an 11- item instrument addressing key domains; the resulting instrument was judged by the expert panel to have face and content validity.[7]

Strengths and weaknesses of the study

This is a prospective external validation study. We compared the new instrument to an independent and reliable gold standard designed for assessing the quality of systematic reviews, allowing multiple testing of convergent validity.

The analytical methods for assessing quality and measuring agreement amongst assessors need further discussion and development. We calculated chance-corrected agreement, using the kappa statistic.[57, 65] While avoiding high levels of agreement due to chance, kappa has its own limitations that have lead to academic criticism.[66, 67] One of the major difficulties with kappa is that when the proportion of positive ratings is extreme, the possible agreement above chance agreement is small and it is difficult to achieve even moderate values of kappa. Thus, if one uses the same raters in a variety of settings, as the proportion of positive ratings becomes extreme, kappa will decrease even if the manner in which the assessors rate the quality does not change. To address this limitation, we also calculated chance-independent agreement using PHIΦ, a relatively new approach to assessing observer agreement.[55, 58]

We were unable to test our convergent validity hypothesis about conflict of interest because of missing data in the systematic reviews and primary studies. This highlights the need for journals and journal editors to require that the information be provided.

Our results are based on a small sample of systematic reviews in a particular clinical area and a relatively small number of AMSTAR assessors. There is a need for replication in larger and different data sets with more diverse appraisers.

Possible mechanisms and implications for clinicians or policymakers

Existing systematic review appraisal instruments did not reflect current evidence on potential sources of bias in systematic reviews and were generally not validated. The best available instrument prior to the development of AMSTAR was OQAQ which was formally validated. However, users of OQAQ frequently had to develop their own rules for operationalizing the instrument and OQAQ does not reflect current evidence on sources of potential bias in systematic reviews (for example funding source and conflict of interest).[68, 69, 70]

Quality assessment instruments can focus on either reporting quality (how well systematic reviewers have reported their methodology and findings (internal validity) or methodological quality (how well the systematic review was conducted (literature searching, pooling of data, etc.). It is possible for a systematic review with poor methodological quality to have good reporting quality. For this reason, the AMSTAR items focus on methodological quality.

Decision-makers have spent the last 10 years trying to work out the best way to use the enormous amounts of systematic reviews available to them. They can hardly know where to start when deciding whether the relevant literature is valid and of the highest quality. AMSTAR is a user friendly methodological quality assessment that has the potential to standardize appraisal of systematic reviews. Early experience suggests that relevant groups are finding the instrument useful.

Unanswered questions and future research

Further validation of AMSTAR is needed to assess its validity, reliability and perceived utility by appraisers and end users of reviews across a broader range of systematic reviews. We need to assess the responsiveness of AMSTAR looking at its sensitivity to discriminate between high and low methodological quality reviews.

We need to assess the applicability of AMSTAR for reviews of observational (diagnostic, etiological and prognostic) studies and if necessary develop AMSTAR extensions for these reviews.

We plan to update AMSTAR as new evidence regarding sources of bias within systematic reviews becomes available.

Acknowledgements

We would like to thank our panel of assessors: Fida Ahmed, Marisol Betancourt, Daniel Francis, David Henry, Avtar Lal, Martin Olmos, Dana Paul, Sumeet Singh and Changhua Yu.

We also thank Dr. Giuseppe G.L. Biondi-Zoccai and Crystal Huntly-Ball for their helpful suggestions on this manuscript.

Figure 1:
Bland and Altman limits of agreement plot for AMSTAR scores

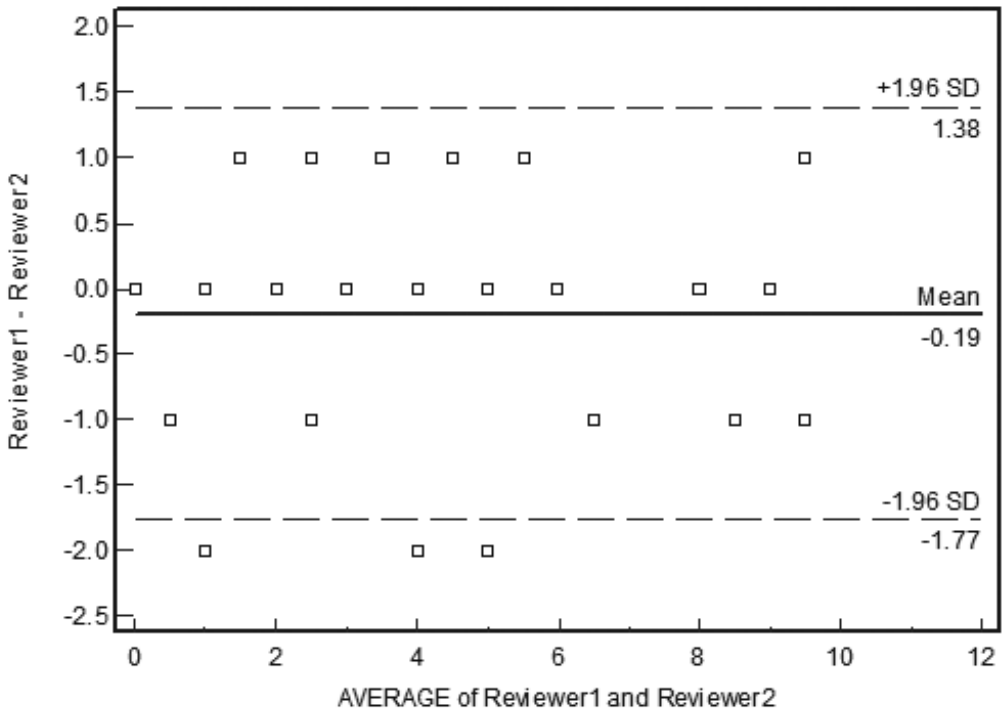


Table 1:
Assessment of the inter-rater agreement for AMSTAR

Items	Kappa (95% CI)	PHI Φ
1. Was an 'a priori' design provided?	0.75 (0.55 to 0.96)	0.76
2. Was there duplicate study selection and data extraction?	0.81 (0.63 to 0.99)	0.83
3. Was a comprehensive literature search performed?	0.88 (0.73 to 1.00)	0.89
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?	0.64 (0.40 to 0.88)	0.64
5. Was a list of studies (included and excluded) provided?	0.84 (0.67 to 1.00)	0.84
6. Were the characteristics of the included studies provided?	0.76 (0.55 to 0.96)	0.76
7. Was the scientific quality of the included studies assessed and documented?	0.90 (0.77 to 1.00)	0.91
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	0.51 (0.25 to 0.78)	0.56
9. Were the methods used to combine the findings of studies appropriate?	0.80 (0.63 to 0.99)	0.80
10. Was the likelihood of publication bias assessed?	0.85 (0.64 to 1.00)	0.85
11. Were potential conflicts of interest included?	1.00 (100% no)	1.00
Overall Score	0.84 (0.67 to 1.00)	0.85

Annex 1: Global assessment rating

Global Assessment rating was assessed using the following instrument.*

How would you rate the scientific quality of the overview?

Extensive Flaws		Major Flaws		Minor Flaws		Minimal Flaws
❒ 1	❒ 2	❒ 3	❒ 4	❒ 5	❒ 6	❒ 7

*Oxman AD, Guyatt GH. (1991) Validation of an index of the quality of review articles. J Clin Epidemiol 44(11): 1271-78.

References

1. Young D. Policymakers, experts review evidence-based medicine. *Am J Health Syst Pharm* 2005; 62(4): 342-343.
2. Dolan-Mullen P, Ramírez G. The Promise and Pitfalls of Systematic Reviews. *Annual Review of Public Health* 2006; 27: 81-102.
3. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44(11): 1271-78.
4. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995; 16(1): 62-73.
5. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. Systematic review in health care meta-analysis in context. London: *BMJ Books* 2001; (7): 122-39.
6. Shea B. Assessing the quality of meta-analyses of randomized controlled trials. MSc thesis. University of Ottawa, Department of Epidemiology and Community Medicine, 1999.
7. Shea B, Grimshaw JM, Wells GA, Boers M, Andersson N, et al. Development of AMSTAR: A Measurement Tool to Assess Systematic Reviews. *BMC Medical Research Methodology* 2007; 7:10.
8. Singh S, Bai A, Lal A, Yu C, Ahmed F, et al. Developing evidence-based best practices for the prescribing and use of proton pump inhibitors in Canada. The Canadian Agency for Drugs and Technologies in Health (CADTH), 2006. Ottawa, Canada.
9. Chiba N, De Gara CJ, Wilkinson JM, Hunt RH. Speed of healing and symptom relief in grade II to IV gastroesophageal reflux disease: a meta-analysis. *Gastroenterology* 1997; 112(6): 1798-810.
10. Caro JJ, Salas M, Ward A. Healing and relapse rates in gastroesophageal reflux disease treated with the newer proton-pump inhibitors lansoprazole, rabeprazole, and pantoprazole compared with omeprazole, ranitidine, and placebo: evidence from randomized clinical trials. *Clin Ther* 2001; 23(7): 998-1017.
11. Klok RM, Postma MJ, van Hout BA, Brouwers JR. Meta-analysis: comparing the efficacy of proton pump inhibitors in short-term use. *Aliment Pharmacol Ther* 2003; 17(10): 1237-45.
12. Van Pinxteren B, Numans ME, Lau J, de Wit NJ, Hungin AP, et al. Short-term treatment of gastroesophageal reflux disease. *J Gen Intern Med* 2003; 18(9): 755-63.
13. Van Pinxteren B, Numans ME, Bonis PA, Lau J. Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease. *Cochrane Database Syst Rev* 2004; (3): CD002095.
14. Rostom A, Dubé C, Wells G, Tugwell P, Welch V, et al. Prevention of NSAID-induced gastroduodenal ulcers. *Cochrane Database Syst Rev* 2002; (4): CD002296.
15. Laheij RJ, van Rossum LG, Jansen JB, Straatman H, Verbeek AL. Evaluation of treatment regimens to cure Helicobacter pylori infection: a meta-analysis. *Aliment Pharmacol Ther* 1999; 13(7): 857-64.
16. Carlsson R, Galmiche JP, Dent J, Lundell L, Frison L. Prognostic factors influencing relapse of oesophagitis during maintenance therapy with antisecretory drugs: a meta-analysis of long-term omeprazole trials. *Aliment Pharmacol Ther* 1997; 11(3): 473-82.
17. Chiba N. Proton pump inhibitors in acute healing and maintenance of erosive or worse esophagitis: a systematic overview. *Can J Gastroenterol* 1997; 11 Suppl B: 66B-73B.
18. Delaney B, Moayyedi P, Deeks J, Innes M, Soo S, et al. The management of dyspepsia: a systematic review. *Health Technol Assess* 2000; 4(39); i,iii-189.
    Available: http://www.ncchta.org/execsumm/summ439.htm.
19. Moayyedi P, Soo S, Deeks J, Delaney B, Harris A, et al. Eradication of Helicobacter pylori for non-ulcer dyspepsia. *Cochrane Database Syst Rev* 2005; (1): CD002096.

20. Moayyedi P, Soo S, Deeks J, Delaney B, Innes M, et al. Pharmacological interventions for non-ulcer dyspepsia. Cochrane Database Syst Rev 2005; (1): CD001960. Available: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD001960/pdf_fs.html (accessed 2006 Feb 16).
21. Delaney BC, Moayyedi P, Forman D. Initial management strategies for dyspepsia. *Cochrane Database Syst Rev* 2003; (2): CD001961.
22. Hopkins RJ, Girardi LS, Turney EA. Relationship between Helicobacter pylori eradication and reduced duodenal and gastric ulcer recurrence: a review. *Gastroenterology* 1996; 110(4): 1244-52.
23. Huang JQ, Sridhar S, Hunt RH. Role of Helicobacter pylori infection and non-steroidal anti-inflammatory drugs in peptic-ulcer disease: a meta-analysis. *Lancet* 2002; 359(9300): 14-22.
24. Moayyedi P, Soo S, Deeks J, Forman D, Mason J, et al. Systematic review and economic evaluation of Helicobacter pylori eradication treatment for non-ulcer dyspepsia. Dyspepsia Review Group. *BMJ* 2000; 321(7262): 659-64.
25. Jovell AJ, Aymerich M, Garcia Altes A, Serra Prat M. Clinical practice guideline for the eradicating therapy of Helicobacter pylori infections associated to duodenal ulcer in primary care. Barcelona: Catalan Agency for Health Technology Assessment, 1998.
26. Gisbert JP, González L, Calvet X, García N, López T. Proton pump inhibitor, clarithromycin and either amoxycillin or nitroimidazole: a meta-analysis of eradication of Helicobacter pylori. *Aliment Pharmacol Ther* 2000; 14(10): 1319-28.
27. Calvet X, García N, López T, Gisbert JP, Gené E, et al. A meta-analysis of short versus long therapy with a proton pump inhibitor, clarithromycin and either metronidazole or amoxycillin for treating Helicobacter pylori infection. *Aliment Pharmacol Ther* 2000; 14(5): 603-09.
28. Gené E, Calvet X, Azagra R, Gisbert JP. Triple vs. quadruple therapy for treating Helicobacter pylori infection: a meta-analysis. *Aliment Pharmacol Ther* 2003; 17(9): 1137-43.
29. Huang J, Hunt RH. The importance of clarithromycin dose in the management of Helicobacter pylori infection: a meta-analysis of triple therapies with a proton pump inhibitor, clarithromycin and amoxycillin or metronidazole. *Aliment Pharmacol Ther* 1999; 13(6): 719-29.
30. Leodolter A, Kulig M, Brasch H, Meyer Sabellek W, Willich SN, et al. A meta-analysis comparing eradication, healing and relapse rates in patients with Helicobacter pylori-associated gastric or duodenal ulcer. *Aliment Pharmacol Ther* 2001; 15(12): 1949-58.
31. Moayyedi P, Murphy B. Helicobacter pylori: a clinical update. *J Appl Microbiol* 2001; (30): 126S-33S.
32. Oderda G, Rapa A, Bona G. A systematic review of Helicobacter pylori eradication treatment schedules in children. *Aliment Pharmacol Ther* 2000; 14(Suppl 3): 59-66.
33. Schmid CH, Whiting G, Cory D, Ross SD, Chalmers TC. Omeprazole plus antibiotics in the eradication of Helicobacter pylori infection: a meta-regression analysis of randomized, controlled trials. *Am J Ther* 1999; 6(1): 25-36.
34. Unge P, Berstad A. Pooled analysis of anti-Helicobacter pylori treatment regimens. *Scand J Gastroenterol* 1996; Suppl 220: 27-40.
35. Unge P. Antimicrobial treatment of H. pylori infection: a pooled efficacy analysis of eradication therapies. *Eur J Surg Suppl* 1998; 582: 16-26.
36. Unge P. What other regimens are under investigation to treat Helicobacter pylori infection? *Gastroenterology* 1997; 113(6 Suppl): S131-S148.
37. Vallve M, Vergara M, Gisbert JP, Calvet X. Single vs. double dose of a proton pump inhibitor in triple therapy for Helicobacter pylori eradication: a meta-analysis. *Aliment Pharmacol Ther* 2002; 16(6): 1149-56.
38. Veldhuyzen van Zanten SJ, Sherman PM. Indications for treatment of Helicobacter pylori infection: a systematic overview. *CMAJ* 1994; 150(2): 189-98.
39. Trépanier EF, Agro K, Holbrook AM, Blackhouse G, Goeree R, et al. Meta-analysis of H pylori (HP) eradication rates in patients with duodenal ulcer (DU). *Can J Clin Pharmacol* 1998; 5(1): 67.

40. Bamberg P, Caswell CM, Frame MH, Lam SK, Wong EC. A meta-analysis comparing the efficacy of omeprazole with H2-receptor antagonists for acute treatment of duodenal ulcer in Asian patients. *J Gastroenterol Hepatol* 1992; 7(6): 577-85.
41. Di Mario F, Battaglia G, Leandro G, Grasso G, Vianello F, et al. Short-term treatment of gastric ulcer: a meta-analytical evaluation of blind trials. *Dig Dis Sci* 1996; 41(6): 1108-31.
42. Eriksson S, Langstrom G, Rikner L, Carlsson R, Naesdal J. Omeprazole and H2-receptor antagonists in the acute treatment of duodenal ulcer, gastric ulcer and reflux oesophagitis: a meta-analysis. *Eur J Gastroenterol Hepatol* 1995; 7(5): 467-75.
43. Poynard T, Lemaire M, Agostini H. Meta-analysis of randomized clinical trials comparing lansoprazole with ranitidine or famotidine in the treatment of acute duodenal ulcer. *Eur J Gastroenterol Hepatol* 1995; 7(7): 661-65.
44. Laine L, Schoenfeld P, Fennerty MB. Therapy for Helicobacter pylori in patients with nonulcer dyspepsia: a meta-analysis of randomized, controlled trials. *Ann Intern Med* 2001; 134(5): 361-9.
45. Mulder CJ, Schipper DL. Omeprazole and ranitidine in duodenal ulcer healing. Analysis of comparative clinical trials. *Scand J Gastroenterol* 1990; Suppl 178: 62-6.
46. Shiau JY, Shukla VK, Dubé C. The efficacy of proton pump inhibitors in adults with functional dyspepsia. Ottawa: Canadian Coordinating Office for Health Technology Assessment, 2002.
47. Danesh J, Lawrence M, Murphy M, Roberts S, Collins R. Systematic review of the epidemiological evidence on Helicobacter pylori infection and non-ulcer or uninvestigated dyspepsia. *Arch Intern Med* 2005; 160(8): 1192-98.
48. Gibson PG, Henry RL, Coughlan JL. Gastro-esophageal reflux treatment for asthma in adults and children. *Cochrane Database Syst Rev* 2005; (3): 1-27.
49. Fischbach LA, Goodman KJ, Feldman M, Aragaki C. Sources of variation of helicobacter pylori treatment success in adults worldwide: a meta-analysis. *Int J Epidemiol* 2002; 31(1): 128-39.
50. Ford A, Delaney B, Moayyedi P. Eradication therapy for peptic ulcer disease in helicobacter pylori positive patients. *Cochrane Database Syst Rev* 2003; (4): CD003840.
51. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44(11): 1271-78.
52. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; 0(1): 622-26.
53. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical Measurement. *Lancet* 1986; i:307-10.
54. Bland JM, Altman DG. Statistical methods for assessing agreement between measurement. *Biochimica Clinic*a 1987; 11: 399-404.
55. Meade M, Cook R, Guyatt G, Groll R, Kachura J, et al. Interobserver Variation in Interpreting Chest Radiographs for the Diagnosis of Acute Respiratory Distress Syndrome. *Am J Respir Crit Care Med* 2000; 161(1); 185-90.
56. Uebersax JS. Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin* 1987; 101: 140-46.
57. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213-220.
58. McGinn T, Guyatt G, Cook R, Meade M. Diagnosis: measuring agreement beyond chance. In: Guyatt G, Rennie D, eds. Users' guide to the medical literature. A manual for evidence-based clinical practice. Chicago, IL: *AMA Press*: 2002; 461-70.
59. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. *BMJ Publishing Group Ltd.* 2005; 330(7499): 1053.
60. Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. *Evaluation & the Health Professions* 2002; 25(1): 116-29.

61. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and Reporting Characteristics of Systematic Reviews. *PLoS Med* 2007; 4(3): e78 doi:10.1371/journal.pmed.0040078.
62. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *New England Journal of Medicine* 1987; 316: 450-54.
63. Bero LA. Managing financial conflicts of interest in research. *Journal of the American College of Dentists* 2005; 72(2): 4-9.
64. Biondi-Zoccai G, Lotrionte M, Abbate A, Testa L. Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. *BMJ* 2006; 332: 202-6.
65. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76: 378-382.
66. McClure M, Willett W. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; 126: 161-169.
67. Cook RJ, Farewell VT. Conditional inference for subject-specific and marginal agreement: two families of agreement measures. *Can J Stat* 1995; 23: 333-344.
68. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* 1998; 279:1566-1570.
69. Cho MK, Bero LA. The quality of drug studies published in symposium proceedings. *Ann Intern Med* 1996; 124:485-489.
70. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003; 326: 1167–1170.

8

8

DISCUSSION

The aim of this research was to explore available measurement instruments for assessing methodological quality and reporting quality, to evaluate the status of these instruments, and to explore the development of a new instrument to fill in any missing gaps. This Chapter provides a summary of the results of all major findings. Next, objectives and a brief discussion of each of the studies and a rationale and discussion of constructive criticisms are discussed. This is followed by recommendations for practice and future directions.

Summary of the major findings for this study

This thesis was comprised of two sets of three related research projects, each guided by its own objective and reported in its own Chapter. The search for instruments to assess quality of systematic reviews (Chapter 2) was necessary to perform the quality assessments of Cochrane reviews described in Chapters 3 and 4, at which point, it became clear that reporting quality (as described in the QUOROM statement) and methodological quality of the review itself were not identical. The former describes how well the reviewers have reported their methodology and findings; the latter considers how well reviewers have conducted their project (literature search, pooling of data, etc.). This seemingly obvious observation complicates discussion about the quality of reviews to date. In addition, our experience with the instrument used in these Chapters, the overview quality assessment questionnaire (OQAQ), demonstrated the instrument needed significant investments in training and time in order to perform adequately, even after we had added descriptors to the items (enhanced OQAQ). Finally, developments in the last 10 years suggested important additional items were needed for a comprehensive assessment of quality.

These considerations lead to the second set of projects (Chapters 5, 6, 7) to develop a new instrument to assess quality of reviews, termed a measurement tool to assess systematic reviews (AMSTAR). A comprehensive list of possible items was drawn up from several instruments and three new items were suggested. This list was reduced and clarified through factor analysis of systematic reviews in which these items had been scored and a subsequent consensus process of experts in a nominal group technique setting. The resulting instrument was validated in a subset of reviews used in the development phase and subsequently in a new set of reviews from a different area of medicine (Chapters 6 and 7).

This thesis brings together a series of studies that documents an incremental process: first, the realization that methodological quality assessment is not the same as reporting quality assessment and that those tools for the former are rarely applied to systematic reviews and are not well validated. Second, that in our view the best tool, the OQAQ, was difficult to apply and needed improvement. Third, that it was better to develop a new tool that would build on the best elements of the OQAQ and the Sacks tool. The development and validation of this instrument, named AMSTAR, for assessing the methodological quality of systematic reviews forms the heart of the thesis.

In the discussion we will summarize and critique each of the Chapters, will bring them up to date with recent developments where possible, and will extract themes that form an agenda for future research.

Objective 1:

To systematically review the instruments available to assess the reporting quality of systematic reviews (Chapter 2)

This review identified three scales (yielding a numerical score) and 21 checklists. Existing tools did not reflect current evidence on systematic reviews and were generally not validated. Most lacked rigour in their development, had large gaps in their accompanying documentation, and were weak in the instructions provided for their use. We found the OQAQ had been rigorously developed[1]. However, we decided to improve the descriptors of the items resulting in the 'enhanced OQAQ' that we have applied in the subsequent Chapters. In addition, but not explicitly mentioned in the published version of the Chapter, the checklist developed by Sacks et al. showed good quality and was especially comprehensive.

Since the publication of this Chapter, there has been a continued proliferation of (largely unvalidated) scales and checklists. For instance, the AHRQ (Agency for Healthcare and Quality) has put out their own guidelines for assessing quality[2]. The University of York has developed a framework for the appraisal of systematic reviews that is being used to enter reviews in the DARE database[3]. These guidelines and framework, however, have not been explicitly validated. In a recent study, the Canadian Agency for Drugs and Technologies in Health (CADTH)[4] searched electronic databases including Medline and the methods database of the Cochrane Library up to and including publications from 2005. They also contacted known methodologists in the field for additional instruments. They identified 52 scales, checklists and guidelines (including the 24 we originally identified) for assessing the quality of systematic reviews.

Objective 2:

To assess the methodological quality and the reporting quality of a complete subset of electronic systematic reviews, by applying OQAQ and QUOROM (Chapter 3)

We applied the enhanced OQAQ and QUOROM to all 57 Cochrane Musculoskeletal (CMSG) systematic reviews published in the Cochrane Database of Systematic Reviews of the Cochrane Library, Issue 4, 2002. Two independent reviewers assessed all reviews with a third reviewer available when needed to help them reach a consensus. They extracted data using prepared forms that included all the items of both instruments. More than half of the CMSG reviews received a rating of 'adequate' on 50% or more of the ten OQAQ quality items. On the overall quality item (range 0-7), the reviews scored relatively well, with only minor flaws identified and a mean score of 5.0 (95% CI: 3.7 to 6.3). On 16 of the 18 QUOROM items, all reviews scored more than 50% of the maximum score.

Thus, the 57 systematic reviews assessed were found to have good overall methodological quality, with scores on individual items revealing only minor flaws.[5] Our study found that the reporting quality of Cochrane musculoskeletal systematic reviews was generally good, although there was room for improvement. This is an important message as there is a widespread perception that since they are Cochrane reviews, they are the best possible quality.

Objective 3:

To determine the impact of updating on the methodological and reporting quality of a subset of systematic reviews (Chapter 4)

We repeated the assessment in Objective 2 with a sample of 53 Cochrane systematic reviews before and after their updating to determine the extent to which their quality had improved. In all cases there was a minimum interval of four months between the publication dates of the original and updated reviews.

In total we assessed 53 systematic reviews. There was no significant improvement with updating in the global quality score of the enhanced OQAQ. The updated reviews did however show a significant improvement on the enhanced OQAQ item assessing whether the conclusions drawn by the author(s) were supported by the data and/or analysis presented in the systematic reviews. But there is clearly still room for improvement of methodological quality.[6] The QUOROM statement showed that the reporting quality of Cochrane reviews improved in some areas with updating, but also leaves room for further improvement.

Currently, Cochrane review updates are carried out primarily to incorporate new findings. Updates should also try to improve any methodological weaknesses of the reviews.

The enhanced OQAQ and QUOROM proved useful in the studies described in Chapters 3 and 4. However, during the course of these studies we had to overcome challenges in their application, even after improving the descriptors of one of them (OQAQ). These challenges included lack of clarity in the questions, lack of published guidance on the application of the quality tools selected for use, the length of time required to apply them, and difficulties with applying their scoring systems. In addition, it became clear that QUOROM explicitly was aimed at reporting quality whereas OQAQ had elements of reporting quality and of methodological quality itself, without distinguishing clearly what was being measured. Subsequently, we formulated the following framework (mentioned already in the Introduction Chapter to orient readers to this distinction): When the quality of a systematic review is examined, two major aspects are assessed. The first, *methodological quality*, considers how well the systematic review was conducted (literature searching, pooling of data, etc.). The second, *reporting quality*, considers how well systematic reviewers have reported their methodology and findings.

Finally, in later years, it was felt important items were missing from these instruments. New evidence had accumulated on the potential for bias in systematic reviews that had not yet been incorporated into a validated instrument. With the above in mind, we decided not to try and improve an existing instrument (OQAQ) further, but to develop a new instrument with improved feasibility, explicitly focused on methodological quality of the review rather than reporting quality.

Objective 4:

To develop a valid and reliable methodological quality assessment instrument for systematic reviews (Chapter 5)

The aim of this objective was to explore the feasibility of developing a generic instrument to assess the methodological quality of systematic reviews. As a first step, a comprehensive list of 37 items was drawn by combining the enhanced OQAQ, the Sacks' instrument and three additional items: language restriction, publication bias, and publication status.

These items were scored in 99 paper-based and 52 electronic systematic reviews and the results obtained were subjected to exploratory factor analysis that identified 11 underlying dimensions (See table 1, Chapter 5). These dimensions were given labels and descriptors. The results of the factor analysis were then subjected to a consensus procedure by nominal group technique. An international panel carried out this consensus exercise, shortening the 37-item assessment tool to a more manageable (11-item) length. The tool was named 'AMSTAR' which is an acronym for a measurement tool to assess systematic reviews.

This approach can be (and was) criticized. The most important points are summarized here (page 108).

The choice of Sacks as a source instrument

We chose the Sacks' instrument out of the list of possible instruments because it was very comprehensive, it came out high in the quality ranking (Chapter 2), and it had been used in a classic survey of methodological quality of systematic reviews over time.[7] The QUOROM statement was not used because it focuses on reporting quality rather than methodological quality, as noted before.

The addition of three items/dimensions

The decision to include and test three additional items was based on emergent evidence about potential sources of bias in systematic reviews. To address a growing consensus among specialists that language was an important methodological issue, we added an additional item on publication language. However, recent evidence has shown that publication language may not be as important an issue as was previously thought.[8] Some empirical evidence indicates that there is no difference in the quality of English and non-English language RCTs.[9,10] Egger et al. found in a retrospective analysis that excluding trials published in languages other than English has generally little effect on summary treatment effect estimates. The importance of non-English language trials is, however, difficult to predict for individual systematic reviews.[11] Language restrictions in systematic reviews have been studied and remain controversial.

Publication bias is the tendency for negative research to get published less frequently, less prominently, more slowly, and the likelihood for positive research to get published more than once. Publication bias has been identified as a major threat to the validity of systematic reviews. Empirical researchers suggest that publication bias is widespread.[12-18] Finally, publication status of studies suggests that published trials are generally larger and may show an overall greater treatment effect than studies published in the 'grey' literature.[19] Grey literature is to be interpreted as those studies not published as a formal journal article (e.g. those found in conference abstracts, books, thesis, government and company reports and other unpublished materials). The importance of including grey literature in all systematic reviews has been discussed.[20] The assessment of the inclusion of grey literature considers whether or not the authors reported searching for grey literature. All three items were subjected to the same appraisal as all the other candidate items in the consensus phase. This resulted in the language item being subsumed under the item on publication status.

Rationale for assessing conflict of interest

The item 'sources of support' was included in the original dataset and came out as a component in the factor analysis. Some may doubt the usefulness of the item concerning conflict of interest. We put a lot of thought into the name and description of this item using all available empirical evidence. In addition to the research previously discussed on this topic, more recent studies also suggest that funding might influence outcomes and quality of research.[21] In this study, the authors concluded that systematic bias favours products which are made by the company funding the research. Explanations included the selection of an inappropriate comparator to the product being investigated and publication bias. In a recently published study by Biondi-Zoccai[22] the authors concluded that reviewers who reported previous not for profit funding were more likely to carry out higher quality systematic reviews. We are convinced that funding sources are associated with bias in systematic reviews. What is most important about this question is that it asks not just about the sources of support for the systematic review itself, but for all primary studies included in the systematic review.

Primary studies may be subject to bias related to the author's competing interests. Djulbegovic et al. found that pharmaceutical industry-sponsored studies were more likely to result in favorable evaluations of new treatments.[23] That studies conducted to support the efficacy of new treatments tend to show more

favorable results is consistent with the drug approval process. Due to the expense, large phase III studies to support regulatory approval will only be conducted if the pharmaceutical company is relatively certain that its new treatment is efficacious. However, this may not be the situation for smaller RCTs where less financial investment is involved.[24, 25]

A study by Cho and Bero has been used to support the potential for conflict of interest based on funding sources. They found that studies published in pharmaceutical company-sponsored symposia proceedings were significantly more likely to favour the new drug of interest than were studies published in peer-reviewed journals.[26]

The rationale for the composition and size of the source dataset of reviews

In Chapter 5, the source of the reviews used in the development of AMSTAR and the rationale for the sample size was not well described. In brief, the original 151 systematic reviews, including 99 paper-based reviews and 52 Cochrane reviews, were obtained from a database developed by Dr. David Moher's team. At the time of retrieving these 99 reviews, we were interested in assessing the quality of systematic reviews over time. An adequate sample size was difficult to calculate due to lack of available data. We conveniently selected reviews from the two time periods 'Early Years' and 'Later Years'. The 52 Cochrane reviews were all included in the Cochrane database of systematic reviews, The Cochrane Library 1996, Issue 3. In this we felt we had a representative sample from the 'universe' of reviews where AMSTAR would be applied.

The choice for factor analysis and a nominal group consensus process to design AMSTAR

Comprehensiveness and feasibility compete to increase responsiveness and decrease the number of items in the instrument. To get to the best compromise between these extremes, we chose two out of a number of possible approaches. Specifically, we chose factor analysis to define a parsimonious set of dimensions from the full set of items and employed an international expert consensus panel to pick one item from each dimension that would be most applicable. The items were subjected to factor analysis and only those items that loaded highly on one component (>.50) were retained. The described factor analysis made it possible to reduce the 37-item instrument to a shorter (29-items) instrument that measured 11 components.

The panel that formed the nominal group comprised eleven experts in the fields of methodological quality assessment and systematic reviews. The group was selected from three organizations involved in both the conduct of systematic reviews and the assessment of methodological quality. It included clinicians, methodologists, epidemiologists and reviewers new to the field. Some, but not all, were previously involved in the Cochrane Collaboration. In their examination on the results of the factor analysis, they reflected critically on the components identified and decided on the items to be included in the new instrument.

The nominal group discussed all 11 components. The items most appropriate for the components were included in the draft instrument. The instrument is an 11-item questionnaire that asks reviewers to answer yes, no, can't answer or not applicable. A separate question on language was identified in the factor analysis as a significant issue, but the nominal group felt that the contradictory evidence in the literature warranted removing this item from the shortened item list and capturing it under the question on publication status.

Objective 5:

To test the validity and reliability of AMSTAR in the source dataset (Chapter 6)

We tested agreement, reliability, construct validity and feasibility (time to complete) of the new instrument. Two assessors independently applied the enhanced OQAQ, the Sacks list, and AMSTAR to a randomly selected sample of 30 reviews out of the 151 systematic reviews used in Chapter 5, which were randomly selected from the Cochrane Library and a database of meta-analysis published in paper-based journals.

The inter-rater reliability of the individual items in the new instrument was substantial, with a mean kappa of 0.70. Kappas recorded for the other instruments were OQAQ 0.63 and Sacks' instrument 0.40. Agreement was lowest on the AMSTAR items assessing status of publication, scientific quality assessment and combinability of statistical data. It should also be noted that overall agreement on these items was good, so the relatively low kappas should be seen as caused by skewness in the responses, i.e. a majority of responses falling into either the 'yes' or the 'no' category. This is a well-known limitation of the kappa statistic.[27, 28]

The construct validity results of the new instrument, expressed as a percentage of the maximum score, showed convergence with the results of the other instruments. Intra-class correlations (ICC) were 0.66 with OQAQ and 0.83 with Sacks. The ICC obtained when comparing OQAQ to Sacks was 0.86. AMSTAR proved highly feasible taking 10-15 minutes to complete compared with OQAQ (taking on average more than 20 minutes), and Sacks (taking on average over 40 minutes). Qualitative analyses of the responses lead to minor changes in the wording of three items. This exercise suggested that the new 11-item instrument has good content and construct validity, good reliability, and excellent feasibility. However, the exercise was limited because it reused the reviews employed for the factor analysis, with the appraisals performed by the developers of the instrument. This could possibly result in overestimation of the reliability and validity of AMSTAR. Our involvement in the development of the instrument could influence how we interpret and apply the criteria that may be different to a novice user. Therefore, a second validation exercise was performed in a new set of reviews and appraised by assessors new to AMSTAR.

Objective 6:

To externally test the validity and reliability of AMSTAR (Chapter 7)

Two 'AMSTAR-naive' assessors applied the instrument to a set of 42 reviews assessing the use of protein pump inhibitors for gastro esophageal reflux disease, dyspepsia and peptic ulcer disease. In the absence of a gold standard, we assessed construct validity by comparing AMSTAR with a validated global scale undertaken by seven assessors with expertise in clinical medicine, epidemiology, measurement, and research methods. This global score is a seven point scale (1 'major flaws', 7 'minor flaws'). In the validation study by Oxman and colleagues, inter-rater reliability was evaluated by assessing the degree to which different individuals agreed on the scientific quality of a set of reports.[29]

The sample of 42 reviews adequately covered a wide range of methodological quality.[30] The inter-observer reliability of the total score was excellent for AMSTAR: kappa 0.84 and Pearson's 0.96. The inter-rater agreement (kappa) between two raters for the global assessment was 0.63. Construct validity was shown by AMSTAR's convergence with the results of the global assessment instrument: Spearman's rank correlation coefficient 0.72.

Both AMSTAR and the global assessment required on average 15 minutes to complete, but with the latter, assessors expressed difficulty in reaching a final decision in the absence of comprehensive guidelines. In contrast, AMSTAR was well received.

The choice for using kappa and PHI statistics

One of the major difficulties with kappa is that when the proportion of positive ratings is extreme, the possible agreement above chance agreement is small and it is difficult to achieve even moderate values of kappa. Thus, if one uses the same raters in a variety of settings, as the proportion of positive ratings becomes extreme, kappa will decrease even if the way the assessors rate the quality does not change. To address this limitation, we also calculated chance-independent agreement using PHI, a relatively new approach to assessing observer agreement.[31-32]

Recommendations for practice and future research directions

The AMSTAR instrument has good inter-observer reliability and validity. The work reported in this thesis offers ample scope for future research. We think we have made a useful contribution to the assessment of the methodological quality of systematic reviews, but like all instruments designed to measure quality, the instrument needs to evolve to keep pace with the new developments in the field it is used. Research into the development of valid and reliable assessment methods will continue to be an integral part of the refinement of the quality of systematic reviews. We want to emphasize a few research priorities specifically.

Further validation and development of AMSTAR for reviews of effectiveness

Further validation is needed to replicate the initial promising validations involving a broader range of assessors and a broader range of reviews to assess whether the reliability and validity are confirmed in diverse circumstances.

Extension of AMSTAR to other types of reviews

Assessing the quality of different types of studies is complex and requires multiple methods of assessment. Further, for the assessment techniques to be useful in a health setting, the procedures and methods need to be practical and easily implemented.

We propose additional steps in the development of AMSTAR: 1) to explore the capacity of AMSTAR to deal with both RCTs and observational studies of therapeutic efficacy, and 2) to develop different AMSTAR versions for observational studies of aetiology, diagnosis or prognosis.

Methodological research on conduct and reporting of systematic reviews

Publication bias remains an area of contention amongst those who assess the quality of systematic reviews. It remains a research priority because it is unclear what the impact of publication bias is on making decisions in health care. We are aware of the passed 20 years of work in this area of research. This has given us some clear answers as to the effect publication bias may have on the overall results of estimating the impact of interventions. This is indeed likely to be the case with techniques to identify and quantify publication bias.[33] Although a number of alternative tests for publication bias exist, none have yet been validated.[34] Inevitably, new evidence will modify current thinking in some areas and at that point, the AMSTAR may be updated.

Detecting differences between the methodological quality and reporting quality: conducting a study to ensure AMSTAR is measuring what it is supposed to measure (i.e. methodological quality). This could be studied in future validation studies.

Promoting Use of AMSTAR

While we are optimizing the validity of AMSTAR, we will need to persuade researchers and decision makers to use the instrument. Our ongoing challenge will be to have AMSTAR widely known. This will involve the development of an implementation strategy.

Our new instrument builds upon previous work completed. Methodologists continue to struggle with methodological quality issues while decision makers struggle with the challenge of basing policy, clinical or resource planning decisions on the available evidence. The personal feedback received on AMSTAR has been supportive. With its publication in a peer review journal[35], we hope it will help many reviewers with their task of assessing the methodological quality and incorporating the results into their systematic reviews.

References

1. Oxman AD. Checklists for reviews articles. *BMJ* 1994 Sep; 309(6955): 648-51.
2. The National Library of Medicine Health Services/Technology Assessment www.ncbi.nlm.nih.gov.
3. The University of York http://www.york.ac.uk/inst/crd/crddatabases.htm.
4. Canadian Agency for drugs and technologies in health www.cadth.ca.
5. Shea B, Bouter L, Grimshaw J, Francis D, Ortiz Z, Wells GA, Tugwell P, Boers M. Scope for Improvement in the Reporting quality of Systematic Reviews from the Cochrane Musculoskeletal Group. *J Rheumatol* 2006; Jan; 33(1): 9-15.
6. Shea B, Boers M, Grimshaw J, Hamel C, Bouter L. Does updating improve the methodological and reporting quality of review quality and the reporting quality of Cochrane reviews? *BMC Medical Research Methodology* 2006; 6: 27.
7. Sacks HS, Berrier J, Reitman D, Anocaon-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *NEJM* 1987; 316: 450-455.
8. Pham B, Lawson M, Klassen T, Moher D. Language of publication restrictions and estimates of an intervention's effectiveness: know your intervention. In Press, *J Clin Epidemiol.*
9. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, Liberati A. What contributions do languages other than English make on the results of meta-analyses? *J Clin Epidemiol* 2000; 53: 964-72.
10. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997; 350: 326-29.
11. Egger M, Juni P, Barlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? *Empirical study Health Technol Assess* 2003; 7(1).
12. Pai M, McCulloch M, Colford JJr, Bero LA. Assessment of Publication Bias in Systematic Reviews on HIV/AIDS. http://www.igh.org/Cochrane/pdfs/MSRI_workshop_talk_abstract.pdf.
13. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; 263: 1385-89.
14. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997; 9: 15-21.
15. Pham B, Platt R, McAuley L, Sampson M, Klassen T, Moher D. Detecting and minimizing publication bias. A systematic review of methods. Technical report; Thomas C. Chalmers Centre for Systematic Reviews, Ottawa, Canada 2000.
16. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997; 315: 640-45.
17. Sterne J.A.C., Gavaghan D, Egger M. Publication and related bias in meta-analysis, power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000; 53: 1119-29.
18. Sutton A, Duval S, Tweedie R, Abrams K, Jones D. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000; 320:1574-77.
19. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. In: *The Cochrane Library*, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.
20. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence the estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000; 356: 1228-31.
21. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* 1998; 279: 1566-1570.
22. Biondi-Zoccai G, Lotrionte M, Abbate A, Testa L. Compliance with QUOROM and reporting quality of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. *BMJ* 2006; 332: 202-6.

23. Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. *Lancet* 2000; 356: 635-638.
24. Dong BJ, Hauck WW, Gambertoglio JG, et al. Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism. *JAMA* 1997; 277: 1205-1213.
25. Rennie D. Thyroid storm. *JAMA* 1997; 277: 1238-1243.
26. Cho MK, Bero LA. The quality of drug studies published in symposium proceedings. *Ann Intern Med* 1996; 124: 485-489.
27. McClure M, Willett W. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; 126: 161-169.
28. Feinstein AR. Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43(6): 543-49.
29. Oxman AD. Checklists for reviews articles. *BMJ* 1994 Sep; 309(6955): 648-51.
30. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Zulma Ortiz, Tim Ramsay, Annie Bai, Vijay K. Shukla, Jeremy M. Grimshaw. External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). PLoS ONE 2007:2(12): e1350. doi:10.1371/journal.pone.0001350.
31. Cook RJ, Farewell VT. Conditional inference for subject-specific and marginal agreement: two families of agreement measures. *Can J Stat* 1995; 23: 333-344.
32. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003; 326: 1167–1170. 10.1136/bmj.326.7400.1167.
33. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ* 2006; 333: 597-600.
34. Rothstein HR, Sutton AJ, Borenstein M, eds. Publication bias in meta-analysis: prevention, assessment and adjustments. Sussex: John Wiley and Sons; 2005.
35. Shea B, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Moher D, Tugwell P, Bouter LM. Development of AMSTAR: A Measurement Tool to Assess Systematic Reviews. *BMC Medical Research Methodology* 2007: 7:10.

Instruments

Instrument 1

Quality of reporting of meta-analyses - QUOROM for clinical randomized controlled trials (RCTs)

Heading	Descriptor	Reported (Y/N)	Page #
Title	Identify the report as a meta-analysis (or systematic review) of randomized trials.		
Abstract	Use a structured format.		
Objectives	The clinical question explicitly.		
Data sources	The databases (i.e., list) and other information sources.		
Review methods	The selection criteria (i.e., population, intervention, outcome, and study design); methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis) in sufficient detail to permit replication.		
Conclusion	The main results.		
Introduction	The explicit clinical problem, biologic rationale for the intervention, and rationale for review.		
Methods			
Searching	The information sources, in detail (e.g., databases, registers, personal files, expert informants, agencies, hand-searching), and any restrictions (years considered, publication status, language of publication.		
Selection	The inclusion and exclusion criteria (defining population, intervention principal outcomes, and study design).		
Validity assessment	The criteria and process used (e.g., masked conditions, quality assessment and their findings).		
Data abstraction	The process used (e.g., completed independently, in duplicate).		
Study characteristics	The type of study design, participants' characteristics, details of intervention, outcome definitions, etc.; and how clinical heterogeneity was assessed.		
Quantitative data synthesis	The principal measures of effect (e.g., relative risk), method of combining results (statistical testing and confidence intervals), handling of missing data, etc.; how statistical heterogeneity was assessed; a rationale for any a priori sensitivity and subgroup analyses; and any assessment of publication bias.		

Instrument 1 continued

Results			
Trial flow	Provide a meta-analysis profile summarizing trial flow (figure 1).		
Study characteristics	Present descriptive data for each trial (e.g., age, sample size, intervention, dose, and duration, follow-up).		
Quantitative data synthesis	Report agreement on the selection and validity assessment; present simple summary results (for each treatment group in each trial, for each primary outcome); data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (e.g., 2 x 2 tables of counts, means and standard deviations, proportions).		
Discussion	Summarize the key findings; discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process (e.g., publication bias); and suggest a future research agenda.		

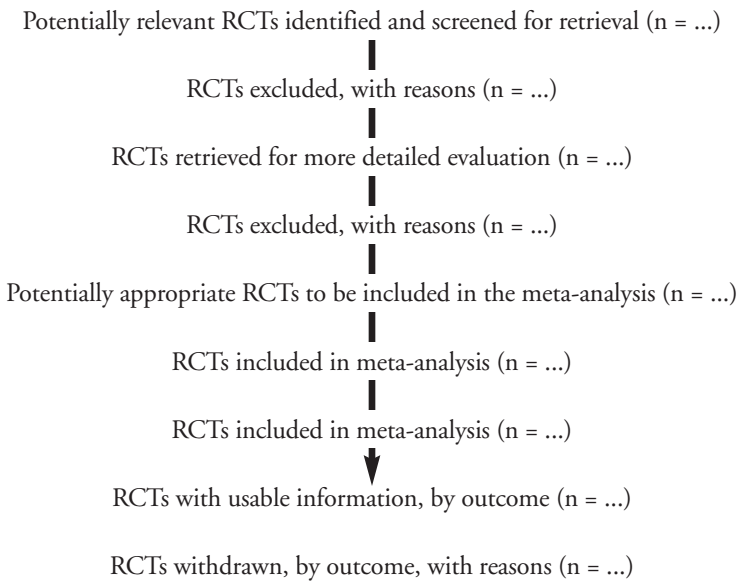


Figure 1

Progress through the stages of a meta-analysis, including selection of potentially relevant randomized controlled trials [RCTs], included and excluded RCTs with a statement of the reasons, RCTs with usable information, and RCTs withdrawn by outcome with a statement of the reasons for the withdrawal.

* Modified from Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354:1896–900.

Instrument 2

Enhanced overview quality assessment questionnaire (OQAQ) originally developed by Oxman, Guyatt et al.

The purpose of this index is to evaluate the scientific quality (i.e. adherence to scientific principles) of research overviews (review articles) published in the medical literature. It is not intended to measure literary quality, importance, relevance, originality, or other attributes of overviews.

The index is for assessing overviews of primary ("original") research on pragmatic questions regarding causation, diagnosis, prognosis, therapy or prevention. A research overview is a survey of research. The same principles that apply to epidemiologic surveys apply to overviews: a question must be clearly specified, a target population identified and accessed, appropriate information obtained from that population in an unbiased fashion, and conclusions derived, sometimes with the help of difference between overviews and epidemiologic surveys the unit of analysis, not the scientific issues that the questions in this index address.

Since most published overviews do not include a methods section, it is difficult to answer some of the questions in the index. Base your answer as much as possible on the information provided in the overview. If the methods that were used are reported incompletely relative to a specific item, score that item as "partially". Similarly, if there is no information provided regarding what was done relative to a particular question, score it as "can't tell" unless there is information in the overview to suggest either that the criterion was or was not met.

1. Were the search methods used to find evidence (original research) on the primary question(s) stated?

❒ yes ❒ partially ❒ no

Yes is given to meta-analysis reporting categories of sources, including years (e.g., databases-Medline) used and whether these categories were named (e.g. Medline). Partial points are given for the category of sources (e.g., electronic, hand, register) named.

2. Was the search for evidence reasonably comprehensive?

❒ yes ❒ can't tell ❒ no

Yes is given if at least three categories, one of which must be electronic with key words stated and any two others (e.g., hand, register) are reported. Key words and/or MESH terms must be stated.

3. Were the criteria used for deciding which studies to include in the overview reported?

❒ yes ❒ partially ❒ no

This item was thought to be reasonably explicit. If 2 or more items mentioned, yes, if <2 mentioned, partially, if none mentioned, no.

4. Was bias in the selection of studies avoided?

❒ yes ❒ can't tell ❒ no

Yes is given if at least two reviewers independently assess for inclusion. A consensus must be reached.

5. Were the criteria used for assessing the validity of the included studies reported?

❒ yes ❒ partially ❒ no

It was felt that the issues relating to publication bias should not be included in the assessment of this. Yes is given to those meta-analysis reporting 'a priori' methods of validity assessment (e.g., if the author(s) chose to include only randomized, double-blind, placebo controlled trials, or allocation concealment as inclusion criteria).

6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?

❒ yes ❒ can't tell ❒ no

This item relates to validity assessment. Yes is given if there is a description of any criteria (either internal or external) used either for inclusion, or for analysis (e.g., sensitivity analysis).

7. Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?

❒ yes ❒ partially ❒ no

This item was thought to be reasonably explicit.

8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?

❒ yes ❒ can't tell ❒ no

For question 8, if no attempt was made to combine findings, and no statement is made regarding the inappropriateness of combining findings, check "no". If a summary (general) estimate is given anywhere in the abstract, the discussion, or the summary section of the paper, and it is not reported how the estimate was derived, mark "no" even if there is a statement regarding the limitations of combing the findings of the studies reviewed. If in doubt mark "can't tell".

9. Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?

❒ yes ❒ partially ❒ no

For an overview to be scored as "yes" on question 9, data (not just citations) must be reported that supports the main conclusions regarding the primary question(s) that the overview addresses. If the overview concerns diagnostic/prognostic tests, 'retest is not required' (this ensures that diagnostic/prognostic papers are not scored more rigorously than clinical papers).

10. How would you rate the scientific quality of the overview?

Extensive Flaws		Major Flaws		Minor Flaws		Minimal Flaws
❒ 1	❒ 2	❒ 3	❒ 4	❒ 5	❒ 6	❒ 7

The score for question 10, the overall scientific quality, should be based on your answers to the first nine questions. The following guidelines can be used to assist with deriving a summary score. If the "can't tell" option is used one or more times on the preceding questions, a review is likely to have minor flaws at best and it is difficult to rule out major flaws (i.e. a score of 4 or lower). If the "no" option is used on question 2, 4, 6 or 8, the review is likely to have major flaws (i.e. a score of 3 or less, depending on the number and degree of the flaws).

Instrument 3
A measurement tool to assess systematic reviews (AMSTAR)

1. Was an 'a priori' design provided? The research question and inclusion criteria should be established before the conduct of the review.	Yes ❒ No ❒ Can't answer ❒ Not applicable
2. Was there duplicate study selection and data extraction? There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.	Yes ❒ No ❒ Can't answer ❒ Not applicable
3. Was a comprehensive literature search performed? At least two electronic sources should be searched. The report must include years and databases used (e.g. Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.	Yes ❒ No ❒ Can't answer ❒ Not applicable
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion? The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.	Yes ❒ No ❒ Can't answer ❒ Not applicable
5. Was a list of studies (included and excluded) provided? A list of included and excluded studies should be provided.	Yes ❒ No ❒ Can't answer ❒ Not applicable
6. Were the characteristics of the included studies provided? In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the studies analyzed e.g. age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.	Yes ❒ No ❒ Can't answer ❒ Not applicable
7. Was the scientific quality of the included studies assessed and documented? A priori' methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.	Yes ❒ No ❒ Can't answer ❒ Not applicable
8. Was the scientific quality of the included studies used appropriately in formulating conclusions? The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.	Yes ❒ No ❒ Can't answer ❒ Not applicable

9. Were the methods used to combine the findings of studies appropriate? For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e. Chi-squared test for homogeneity, I2). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e. is it sensible to combine?).	Yes ❒ No ❒ Can't answer ❒ Not applicable
10. Was the likelihood of publication bias assessed? An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test).	Yes ❒ No ❒ Can't answer ❒ Not applicable
11. Was the conflict of interest included? Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.	Yes ❒ No ❒ Can't answer ❒ Not applicable

Shea et al. BMC Medical Research Methodology 2007 7:10 doi:10.1186/1471-2288-7-10

Summary

Systematic reviews have become the standard method for summarizing data regarding the effects of healthcare interventions. In the last decade there has been a remarkable proliferation of systematic reviews as one of the key tools for evidence-based health care. Systematic reviews within health care are usually conducted retrospectively and are susceptible to a range of potential sources of bias. However, there has been no agreement on the best approach to assessing their methodological quality. The aim of this thesis is to pull together the best available instruments to develop, update and validate them, and to produce a reliable, practical and convenient tool that can be used in a variety of settings.

To achieve these aims we carried out a series of interlinked studies.

In Chapter 1 we provide a brief summary of the existing literature on assessing the methodological and reporting quality of systematic reviews. We list our objectives, along with the measurement instruments and systematic reviews studied in this thesis.

In Chapter 2 we undertook a systematic review of available instruments used to assess the quality of systematic reviews. We compiled and appraised a complete list of all available tools for the assessment of systematic reviews. We described the development of the QUOROM (quality of reporting of meta-analysis) statement and compared it to other instruments identified through a systematic review. Finally, we improved the descriptors of the instrument that rated highest (the overview quality assessment questionnaire-OQAQ). We found that the literature described many checklists and scales for use as evaluation tools, but most were missing important evidence-based items. A pilot study suggested considerable room for improvement in the reporting of systematic reviews using different instruments.

In Chapter 3 we assessed the methodological and reporting quality of a set of systematic reviews by applying OQAQ and QUOROM to all 57 Cochrane musculoskeletal systematic reviews published in the Cochrane Database of Systematic Reviews of the Cochrane Library. We found good overall methodological quality, with scores on individual items revealing only minor flaws. However, we concluded more work was needed in reporting search results, documentation of the flow of studies, identification of the type of studies, summary of the key findings and the need for specific guidelines for reporting protocols.

In Chapter 4 we determined the impact of updating on the methodological quality and reporting quality of a set of systematic reviews. Under this objective we assessed a newly selected sample of systematic reviews before and after their updating using the same two instruments. The sample covered a wide variety of health topics published in the Cochrane Library. We assessed the updated and original versions of the systematic reviews using two instruments: the 10 item OQAQ, and the 18-item QUOROM statement. In total, 53 systematic reviews were evaluated. Updating produced no significant improvement in the global quality score of the OQAQ. Updated reviews showed a significant improvement on the OQAQ item assessing whether the conclusions drawn by the author(s) were supported by the data and /or analysis presented in the systematic review. The QUOROM data confirmed these findings. After the studies in Chapters 3 and 4 we concluded that there was room for a new instrument focused on methodological quality (rather than reporting quality) of systematic reviews, with improved content and feasibility.

In Chapter 5 we developed such an instrument by building upon previous tools, empirical evidence and expert consensus. To start, a 37-item assessment tool was formed by combining two instruments developed to assess methodological quality 1) the enhanced OQAQ, 2) a checklist created by Sacks et al., and 3) three additional items recently judged to be of methodological importance. This tool was applied to 99 paper-based and 52 electronic systematic reviews. Exploratory factor analysis identified 11 underlying components. From each component, methodological experts selected one item through a nominal group consensus process to arrive at a feasible assessment tool with face and content validity. The 11-item tool was named AMSTAR: A MeaSurement Tool to Assess systematic Reviews.

In Chapter 6 we tested the construct validity, reliability and feasibility of AMSTAR in the source dataset. We tested the new instrument by having two assessors apply it, as well as the two original instruments, to a random sample of 30 systematic reviews (out of the 151 selected for Chapter 5). The construct validity results of the new instrument, expressed as a percentage of the maximum score, showed convergence with the results of the other instruments. Intra-class correlations (ICC) were 0.66 with OQAQ and 0.83 with Sacks' checklist. The ICC obtained when comparing OQAQ to Sacks' checklist was 0.86. AMSTAR proved highly feasible taking 10-15 minutes to complete compared with OQAQ (taking on average more than 20 minutes), and Sacks' checklist (taking on average over 40 minutes). Qualitative analyses of the responses lead to minor changes in the wording of three items. This internal validation exercise suggested that the new 11-item instrument has good content and construct validity, good reliability, and excellent feasibility.

In Chapter 7 we further tested the reliability and external validity of AMSTAR using a separate set of reviews. External assessors, with no prior exposure to AMSTAR, applied the instrument to a set of 42 systematic reviews focusing on therapies to treat gastro-esophageal reflux disease, peptic ulcer disease, and other acid-related diseases. In the absence of a gold standard, we assessed construct validity by comparing AMSTAR with a global scale undertaken by seven assessors with expertise in clinical medicine, epidemiology, measurement, and research methods. The inter-observer reliability of the total score was excellent for AMSTAR: kappa 0.84 and Pearson's 0.96. The inter-rater agreement (kappa) between two raters for the global assessment was 0.63. Construct validity was shown by AMSTAR's convergence with the results of the global assessment instrument. Both AMSTAR and the global assessment required on average 15 minutes to complete, but with the latter, assessors expressed difficulty in reaching a final decision in the absence of comprehensive guidelines. In contrast, AMSTAR was well received.

Conclusions

The aim of the research reported in this thesis was to explore available measurement instruments for assessing methodological quality and reporting quality, to evaluate the status of these instruments, and to explore the development of a new instrument to fill in any missing gaps.

Assessing the methodological quality of different types of studies is complex and requires multiple methods of assessment. Further, for the assessment techniques to be useful in a health setting, the procedures and methods need to be practical and easily implemented. While we are planning to perform additional validation the real test will be to persuade researchers and decision makers to use the instrument. Our ongoing challenge will be to have AMSTAR widely used. This will involve the development of an implementation strategy. The early signs are good - AMSTAR has been adopted or recommended by a number of groups, including the Canadian Agency for Drugs and Technologies in Health.

Methodologists continue to struggle with methodological quality issues while decision makers struggle with the challenge of basing policy, clinical or resource planning decisions on the available evidence. The personal feedback received on AMSTAR to date has been supportive. With its publication in peer reviewed journals and in this thesis, we hope that it will help many reviewers with their tasks of assessing the methodological quality of systematic reviews and incorporating their results in clinical and policy decisions.

Beoordeling van de methodologische kwaliteit van systematische reviews
De ontwikkeling van AMSTAR

Systematische reviews zijn de standaardmethode geworden voor het samenvatten van gegevens met betrekking tot de effecten van interventies in de gezondheidszorg, en worden als een van de belangrijkste instrumenten voor 'evidence-based healthcare' beschouwd. In het afgelopen decennium is het aantal systematische reviews aanzienlijk toegenomen. Systematische reviews worden retrospectief uitgevoerd en zijn daardoor gevoelig voor verschillende mogelijke bronnen van vertekening. Er is echter geen overeenstemming over de beste benadering voor het beoordelen van de methodologische kwaliteit van systematische reviews. Het doel van dit proefschrift is het vergaren van de beste beschikbare instrumenten, het ontwikkelen, aanpassen en valideren van die instrumenten en het produceren van een betrouwbaar, praktisch en werkbaar instrument dat in verschillende settings kan worden gebruikt.

Om deze doelen te bereiken, hebben we een aantal onderling samenhangende studies uitgevoerd.

In hoofdstuk 1 geven we een kort overzicht van de bestaande literatuur over het beoordelen van de kwaliteit van de methodologie en van de rapportage van systematische reviews. We beschrijven onze vraagstellingen en de meetinstrumenten en systematische reviews die in dit proefschrift zijn bestudeerd.

In hoofdstuk 2 hebben we een systematisch onderzoek uitgevoerd naar de beschikbare instrumenten die worden gebruikt voor het beoordelen van de kwaliteit van systematische reviews. We hebben een volledige lijst samengesteld en alle beschikbare instrumenten voor de beoordeling van systematische reviews geëvalueerd. We beschreven de ontwikkeling van het QUOROM-statement (Quality Of Reporting Of Meta-analysis) en vergeleken dit met andere instrumenten. Ten slotte hebben we de descriptoren van het instrument met het beste resultaat (de OQAQ, Overview Quality Assessment Questionnaire) verbeterd. Uit ons onderzoek bleek dat in de literatuur veel checklists en schalen worden beschreven die als evaluatie-instrument worden gebruikt, maar dat bij de meeste daarvan 'evidence-based' items ontbreken. Er leek aanzienlijke ruimte te zijn om de beoordeling van de kwaliteit van systematische reviews te verbeteren.

In hoofdstuk 3 beschrijft de beoordeling van de kwaliteit van methodologie en rapportage van een aantal systematische reviews, door OQAQ en QUOROM toe te passen op alle 57 systematische reviews van de Cochrane Musculoskeletal Group (gepubliceerd in de Cochrane Database of Systematic Reviews van de Cochrane Library). In het algemeen was de methodologische kwaliteit goed, met slechts kleine minpunten in de afzonderlijke items. Toch bleek ook dat verdere verbetering mogelijk is bij rapportage van zoekresultaten, documenteren van de onderzoeksstroom, identificeren van het type onderzoek en samenvatten van de belangrijkste resultaten. Tevens bleek er een behoefte aan specifieke richtlijnen voor rapportage.

In hoofdstuk 4 hebben we bekeken of herzieninging ('update') van een review de kwaliteit van methodologie en rapportage verbetert. In een steekproef hebben we een aantal reviews vóór en na de herziening beoordeeld. De steekproef van 53 systematische reviews bestreek een groot aantal verschillende gezondheidswetenschappelijke onderwerpen in de Cochrane Library. Net als in hoofdstuk 3 gebruikten we de OQAQ en de QUOROM voor de beoordeling. Herzieningen gaven geen significante verbetering in de globale kwaliteitsscore van de OQAQ, maar wel op het OQAQ-item waarmee wordt beoordeeld of de getrokken conclusies worden ondersteund door de gepresenteerde gegevens en/of analyse. De QUOROM-gegevens bevestigden deze resultaten. Na de studies in hoofdstuk 3 en 4 concludeerden we dat er ruimte is voor een nieuw instrument gericht op de methodologische kwaliteit (in tegenstelling tot de kwaliteit van de rapportage) van systematische reviews, met een verbeterde inhoud en haalbaarheid.

In hoofdstuk 5 hebben we een dergelijk instrument ontwikkeld door voort te bouwen op eerdere instrumenten, empirisch bewijs en consensus onder experts. Om te beginnen werd een uit 37 items bestaand beoordelingsinstrument gevormd door het combineren van twee instrumenten die zijn ontwikkeld voor de beoordeling van methodologische kwaliteit 1) de verbeterde OQAQ, 2) een door Sacks et al. gemaakte checklist en 3) drie aanvullende items die van methodologisch belang worden geacht. Dit instrument is toegepast op 99 papieren en 52 elektronische systematische reviews. Uit de exploratieve verkennende factoranalyse kwamen 11 onderliggende componenten naar voren. Uit elke component werd door methodologische experts één item geselecteerd via een consensus procedure ('nominal groups'), zodat een werkbaar beoordelingsinstrument met validiteit op het eerste gezicht en op inhoud (face en content validity) kon ontstaan. Het uit 11 items bestaande instrument kreeg de naam AMSTAR: 'A MeaSurement Tool to Assess systematic Reviews' (een meetinstrument voor de beoordeling van systematische reviews).

In hoofdstuk 6 hebben we de constructvaliditeit, betrouwbaarheid en werkbaarheid van AMSTAR getest in de brongegevens-set. We hebben het nieuwe instrument getest door het evenals de twee oorspronkelijke instrumenten door twee beoordelaars te laten toepassen op een willekeurige steekproef van 30 systematische onderzoeken (uit de 151 die waren geselecteerd voor hoofdstuk 5). De resultaten voor de constructvaliditeit van het nieuwe instrument, uitgedrukt als een percentage van de maximumscore, vertoonden convergentie met de resultaten van de andere instrumenten. De intraclass correlatie (ICC) coëfficiënt was 0,66 met OQAQ en 0,83 met de checklist van Sacks. De ICC bij de vergelijking van OQAQ met de checklist van Sacks was 0,86. AMSTAR bleek zeer werkbaar te zijn met een benodigde tijd voor het uitvoeren van 10-15 minuten in vergelijking met OQAQ (gemiddeld meer dan 20 minuten nodig) en de checklist van Sacks (gemiddeld meer dan 40 minuten nodig). Kwalitatieve analyse van de respons leidde tot minimale wijzigingen in de verwoording van drie items. Deze interne valideringstudie suggereerde dat het nieuwe uit 11 items bestaande instrument een goede inhouds- en constructvaliditeit, een goede betrouwbaarheid en een uitstekende werkbaarheid heeft.

In hoofdstuk 7 hebben we de betrouwbaarheid en externe validiteit van AMSTAR verder getest met een afzonderlijke set systematische reviews. Externe beoordelaars zonder ervaring met AMSTAR pasten het instrument toe op een set van 42 systematische reviews die waren gericht op de behandeling van gastro-oesophageale refluxziekte, maagzweren en andere aan maagzuur gerelateerde aandoeningen. Bij afwezigheid van een gouden standaard hebben we de constructvaliditeit beoordeeld door AMSTAR te vergelijken met beoordelingen op een globale schaal die is gehanteerd door zeven beoordelaars met ervaring in klinische geneeskunde, epidemiologie, metingen en onderzoeksmethoden.
De interbeoordelaarsbetrouwbaarheid van de totale score voor AMSTAR was uitstekend: kappa 0,84 en Pearson 0,96. De interbeoordelaarsovereenstemming (kappa) tussen twee beoordelaars voor de globale schaal was 0,63. De constructvaliditeit werd aangetoond door de convergentie van AMSTAR met de resultaten van het globale beoordelingsinstrument. Voor het uitvoeren van zowel AMSTAR als de globale beoordeling was gemiddeld 15 minuten nodig, maar bij de laatste ondervonden beoordelaars moeilijkheden bij het bereiken van een definitieve beslissing, door de afwezigheid van uitgebreide richtlijnen. AMSTAR werd daarentegen goed ontvangen.

Het doel van de studies waarover in dit proefschrift wordt gerapporteerd, was het verkennen van de beschikbare meetinstrumenten voor het beoordelen van de kwaliteit van de methodologie en van de rapportage van systematische reviews, het evalueren van de status van deze instrumenten en het verkennen van de ontwikkeling van een nieuw instrument om eventuele lacunes te vullen.

De beoordeling van de methodologische kwaliteit van verschillende typen systematische reviews is een complexe taak waarvoor meerdere beoordelingsmethoden vereist zijn. Bovendien kunnen de beoordelingstechnieken alleen bruikbaar zijn voor de gezondheidszorg als de procedures en methoden

praktisch en eenvoudig te implementeren zijn. Hoewel we nog van plan zijn een aanvullende validatie uit te voeren, zal de echte test eruit bestaan onderzoekers en besluitvormers over te halen het instrument te gaan gebruiken. Onze voortdurende uitdaging is AMSTAR breed gebruikt te laten worden. Hiervoor moet een implementatiestrategie worden ontwikkeld. De eerste tekenen zijn gunstig. AMSTAR is inmiddels in gebruik genomen of aanbevolen door een aantal groepen, waaronder de Canadian Agency for Drugs and Technologies in Health.

Methodologen blijven worstelen met problemen omtrent de methodologische kwaliteit, terwijl besluitvormers worden uitgedaagd om besluiten op het gebied van beleid, klinische zaken en resourceplanning te baseren op het beschikbare bewijs. De feedback die tot op heden is ontvangen over AMSTAR, was bemoedigend. We hopen dat de publicatie in peer-reviewed tijdschriften en in dit proefschrift veel beoordelaars zal helpen bij de taak om de methodologische kwaliteit van systematisch reviews te beoordelen en hun resultaten mee te nemen in klinische en beleidsbesluiten.

Acknowledgements

ACKNOWLEDGEMENTS

In case this is the only part of my thesis that you will read, I want to ensure you are entertained and impressed by its inclusiveness and internationality. This will become clear as I list the many individuals who have helped me in the last few years.

I want to start with the photographic theme of my thesis - bridges. A bridge is typically thought of as a physical structure that allows people to cross an obstacle. The word 'bridge' has many other meanings. It can be the superstructure of a ship from which it is steered; it is the part of an instrument that transmits the vibrations of the strings to the resonating body; and it is a card game. But it can also be a set of concepts that resemble a bridge in form or function.

For me this thesis has been a bridge. First and foremost it has allowed me to cross an obstacle - the need for a doctorate! More importantly, it has allowed me to bridge concepts - those espoused by others in relation to the quality of systematic reviews of studies of healthcare and health policy interventions. Through bridging them I hope that I have made a significant contribution to this large and expanding field. Lastly, the thesis has enabled me to build bridges with valued colleagues around the world, from whom I have learned so much.

Thanking the many individuals who helped me achieve my dream is a pleasant task, but in compiling this list of friends and colleagues who have helped along the way, I worry about leaving someone out. Forgive me in advance if I have overlooked anyone.

Let me start with my family and friends in North America. They have been extremely patient throughout this journey. When I proposed doing a PhD in Amsterdam they first thought I was losing my mind. Amsterdam is a wonderful academic centre, but has other sides to its character. Would I just be hanging out in 'coffee shops', or would I really be studying and working hard? My family encouraged me to let nothing stand in the way of furthering my education believing that higher education is the key to a happy and healthy family, and life! I'm forever in their debt. Julia, Brian and mom, thank you for making this happen!

My Netherlands family and friends, who took me under their wings, owed me nothing but gave me absolutely everything. I so enjoyed my trips to Mijdrecht and Amsterdam. I will miss spending time studying at the EMGO. Pieter, Tineke and Menno Koopmans, I owe you more than you could ever imagine. Tineke Rotenberg, thank you for allowing them to adopt me into their family! And to Pieter Sipkema for making sure I arrived at the university safely and on time!

My co-promoters, Lex Bouter, Maarten Boers and Jeremy Grimshaw were unbelievable throughout this entire journey. They were there to guide, facilitate and direct every step of the process. It was a task with little return for them. My mentors set no limits and for their ongoing advice and guidance I will be forever grateful. Thank you to Ian Graham, David Henry, David Moher, Zulma Ortiz and Andy Oxman!

I want to give a special mention and recognition for her contribution to this work to my longtime colleague and friend, Candyce Hamel. She endlessly and unselfishly provided her knowledge and assistance to the projects; and to Ellen Visser who made it possible to complete a PhD in a foreign country. As my paranymph, she has been inundated with the task of planning my visits and the formal promotion. Thank you!

In a traditional CIET environment, Neil Andersson provided the support, space and time needed to complete the task. On many occasions he offered these words of guidance "follow your mentors advice and simply get on with it." Thank you Neil!

As I put my final words and thoughts on paper let me repeat my sincere thanks to everyone! If at times I have seemed at all unappreciative, let me assure all of you, that in my heart there is a very big space for everyone in my life who helped make this dream a reality.

Thank you to my examiners, Rob de Bie, Rob Scholten, Peter Tugwell, Riekie de Vet, George Wells, Danielle van der Windt and Bernard Uitdehaag, whose helpful comments helped shape this thesis and finally, to all those who contributed to the science including:

Co-promotors: Maarten Boers, Lex Bouter and Jeremy Grimshaw.

Co-authors: Neil Andersson, Annie Bai, Maarten Boers, Lex Bouter, Catherine Dubé, Daniel Francis, Candyce Hamel, David Henry, Jeremy Grimshaw, Betsy Kristjansson, David Moher, Graham Mowatt, Zulma Ortiz, Joan Peterson, Ashley Porter, Tim Ramsay,Vijay Shukla, Peter Tugwell and George Wells.

Nominal group and panel members: Fida Ahmed, Marisol Betancourt, Maarten Boers, Daniel Francis, David Henry, Tara Horvath, Gail Kennedy, Avtar Lal, Kirby Lee, Kathryn McDonald, Martin Olmos, Zulma Ortiz, Joan Peterson, Madhukar Pai, George Rutherford, Kaveh Shojania, Sumeet Singh, Maria Suarez-Almazor, Peter Tugwell, Vivian Welch, George Wells and Changhua Yu.

Partners: Institute for Research in Extramural Medicine (EMGO Institute) of the VU University Medical Center (VUmc), and the Department of Clinical Epidemiology and Biostatistics (KEB) of the VU University Medical Center, the Netherlands, Community Information and Epidemiological Technologies (CIETcanada), Institute for Population Health, University of Ottawa, Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, Cochrane HIV/AIDS Group, San Francisco and the Cochrane Musculoskeletal Group, Ottawa, Canada.

And for their technical assistance with the production of this book: Julia Donahue, Brian Donahue, Blair Dunne, Peter Frie, Ron Habinski, Candyce Hamel, Keri Hamel, Thelma Hasson, Kylie Hugo, Crystal Huntly-Ball, Tineke & Pieter Koopmans, Jorge Laucirica, Steve Mitchell, Dana Paul and Adam Shea.

DANKWOORD

Het is mogelijk dat dit het enige gedeelte van mijn proefschrift is dat u leest. In dat geval wil ik mijn best doen ervoor te zorgen dat u zich amuseert en dat u onder de indruk raakt van de volledigheid en het internationale karakter. Dat zal duidelijk worden bij de opsomming van de vele personen die me de afgelopen paar jaar hebben geholpen.

Maar ik wil beginnen met het fotografische thema van mijn proefschrift: bruggen. Een brug wordt meestal gezien als een fysiek bouwwerk waarmee mensen een obstakel kunnen oversteken. Maar het woord 'brug' heeft veel andere betekenissen. Het kan het bovenste gedeelte van een schip zijn van waaruit het wordt bestuurd, het is het deel van een instrument waarmee de trillingen van de snaren naar het resonerende lichaam worden overgebracht en het Engelse woord 'bridge' is een kaartspel. Een brug kan ook een aantal concepten zijn die in vorm of functie op een brug lijken.

Voor mij is dit proefschrift een brug. Allereerst heb ik er een obstakel mee kunnen oversteken: de behoefte aan een doctoraat. Maar belangrijker is dat ik er concepten mee heb overbrugd: concepten die door anderen worden aangenomen met betrekking tot de kwaliteit van systematisch onderzoek naar onderzoeken op het gebied van gezondheidszorg en medisch beleid. Door het overbruggen van die concepten hoop ik een duidelijke bijdrage te hebben geleverd aan dit omvangrijke en steeds groter wordende gebied. Ten slotte heb ik via dit proefschrift bruggen kunnen bouwen naar gewaardeerde collega's in de hele wereld, van wie ik veel heb geleerd.

Het bedanken van de vele personen die me hebben geholpen mijn droom te bereiken is een aangename taak, maar bij het samenstellen van deze lijst van vrienden en collega's die me onderweg hebben geholpen, ben ik bang om iemand te vergeten. Vergeef me op voorhand als ik iemand over het hoofd heb gezien.

Ik wil beginnen met mijn familie en vrienden in Noord-Amerika. Zij hebben gedurende deze reis zeer veel geduld getoond. Toen ik voorstelde om mijn PhD in Amsterdam te gaan doen, dachten ze eerst dat ik gek was geworden. Amsterdam is een geweldige academische stad, maar heeft ook andere kanten. Zou ik alleen maar rondhangen in coffeeshops, of zou ik echt hard gaan studeren en werken? Mijn familie heeft me gestimuleerd om niets in de weg te laten staan van het vervolgen van mijn opleiding. Ze geloven dat onderwijs de sleutel is voor een gelukkige en gezonde familie en een gelukkig en gezond leven! Ik ben ze voor eeuwig dankbaar. Julia, Brian en mama: bedankt dat jullie dit mogelijk hebben gemaakt.

Mijn familie en vrienden in Nederland, die me onder hun hoede hebben genomen, waren me niets verschuldigd maar hebben me alles gegeven. Ik heb enorm genoten van mijn reizen naar Mijdrecht en Amsterdam. Ik zal het studeren aan het EMGO Instituut missen.

Pieter, Tineke en Menno Koopmans: aan jullie ben ik meer verschuldigd dan ik kan zeggen. Tineke Rotenberg: bedankt dat je me door hun gezin hebt laten adopteren! En dank aan Peter Sipkema, die ervoor heeft gezorgd dat ik veilig en op tijd aankwam op de universiteit.

Mijn co-promotoren Lex Bouter, Maarten Boers en Jeremy Grimshaw zijn van ongelooflijk belang geweest tijdens deze hele reis. Ze stonden altijd klaar om elke stap van het proces te begeleiden, mogelijk te maken en te regisseren. Dat was een taak waarvoor ze weinig hebben teruggekregen. Mijn mentoren kenden geen grenzen en ik ben voor eeuwig dankbaar voor hun constante advies en sturing. David Henry, Joe Losos en Andy Oxman: bedankt!

Een speciaal woord van erkenning voor haar bijdrage aan dit werk gaat uit naar Candyce Hamel, sinds lange tijd een collega en vriendin. Ze heeft eindeloos en belangeloos met haar kennis en hulp bijgedragen aan de projecten. Ook Ellen Visser, die het mogelijk heeft gemaakt een PhD te voltooien in het buitenland, verdient dank. Als mijn paranimf had ze haar handen vol aan het plannen van mijn bezoeken en de formele promotie. Dankjewel!

In de traditionele CIET-omgeving heeft Neil Andersson me de ondersteuning, ruimte en tijd geboden die ik nodig had om mijn taak te volbrengen en hij heeft me regelmatig aangespoord met de woorden 'gewoon doorgaan'. Bedankt, Neil!

Met deze laatste woorden en gedachten die ik op papier zet, wil ik mijn gemeende dank aan iedereen herhalen. Ik heb soms misschien ondankbaar geleken, maar ik verzeker jullie allemaal dat er in mijn hart een grote plek is voor iedereen in mijn leven die heeft geholpen om deze droom te laten uitkomen.

The author

Born in St. John's Newfoundland, commonly called 'the Rock', I grew up in a house with my three brothers Chris, Terry and Eugene and 25 of my extended Irish family. It was most likely the positive insanity of that environment that instilled the confidence, resilience and persistence needed to take this on.

The rest of my early years were spent with my mother's side of the family from a community of 250 members. I was encouraged to pursue higher levels of education and to follow in my grandmother's footsteps. She was one of the first formally educated teachers living in an outport, traditionally known as a temporary port used by Newfoundland "shore" fishermen.

I give most of the credit for my education to my late father, whose words to me were 'you can do and have anything you want'. Little did I know what that meant at the time! I am also indebted to my mother, who persuaded me to pursue a career in health care. I graduated with a diploma and Follow-up degree in Nursing and moved to the west coast of Canada, but eventually settled in Ottawa, the 'Nations Capital'. Intrigued by health research, I was encouraged to join the Cochrane Collaboration. I owe a great deal to the members of the Collaboration, and all those who preceded me in the field of quality assessment of systematic reviews.

It was the Ottawa-Amsterdam Cochrane connections made while working with the Collaboration and completing a master's in epidemiology, that led me to Amsterdam to pursue a PhD. I am and will be forever grateful!

Bibliography

2115 Citations

Related to the thesis topic:

1. Shea BJ, Hamel C, Bouter LM, Grimshaw J, Kristjansson B, Henry D, Boers M. Internal validation of AMSTAR: a measurement tool to assess systematic reviews (Journal of Clinical Epidemiology 2008).
2. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Zulma Ortiz, Tim Ramsay, Annie Bai, Vijay K. Shukla, Jeremy M. Grimshaw. External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). PLoS ONE 2007:2(12): e1350. doi:10.1371/journal.pone. 0001350.
3. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. BMC Medical Research Methodology 2007;7.
4. Shea B, Bouter LM, Grimshaw JM, Francis D, Ortiz Z, Wells GA, Tugwell PS, Boers M. Scope for improvement in the quality of reporting of systematic reviews. From the Cochrane musculoskeletal group. Journal of Rheumatology 2006;33(1):9-15.
5. Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC Medical Research Methodology 2006;6.
6. Shea B, Dube C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. Systematic Review in Health Care Meta-analysis in context. *BMJ Books* 2001: 122-39.

Additional:

7. Wells GA, Cranney A, Peterson J, Boucher M, Shea B, Robinson V, Coyle D, Tugwell P. Alendronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database of Systematic Reviews (Online) 2008(1).
8. Wells G, Cranney A, Peterson J, Boucher M, Shea B, Robinson V, Coyle D, Tugwell P. Risedronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database of Systematic Reviews (Online) 2008(1).
9. Wells GA, Cranney A, Peterson J, Boucher M, Shea B, Robinson V, Coyle D, Tugwell P. Etidronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database of Systematic Reviews (Online) 2008(1).
10. Kristjansson EA, Robinson V, Petticrew M, MacDonald B, Krasevec J, Janzen L, Greenhalgh T, Wells G, MacGowan J, Farmer A, Shea BJ, Mayhew A, Tugwell P. School feeding for improving the physical and psychosocial health of disadvantaged elementary school children. Cochrane Database of Systematic Reviews (Online) 2007(1).
11. Tugwell P, O'Connor A, Andersson N, Mhatre S, Kristjansson E, Jacobsen MJ, Robinson V, Hatcher-Roberts J, Shea B, Francis D, Beardmore J, Wells GA, Losos J. Reduction of inequalities in health: Assessing evidence-based tools. International Journal for Equity in Health 2006;5.
12. Robinson V, Boers M, Brooks P, Francis D, Judd M, McGowan J, Shea B, Simon LS, Strand V, Tugwell P, Wells GA. Patient-reported pain is central to OMERACT rheumatology core measurement sets. Drug Information Journal 2006;40(1):111-6.
13. Brosseau L, Robinson V, Wells G, Debie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (classes I, II and III) for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online) 2005(4).

14. Khadilkar A, Milne S, Brosseau L, Wells G, Tugwell P, Robinson V, Shea B, Saginur M. Transcutaneous electrical nerve stimulation for the treatment of chronic low back pain: A systematic review. Spine 2005;30(23):2657-66.
15. Shea B, Santesso N, Qualman A, Heiberg T, Leong A, Judd M, Robinson V, Wells G, Tugwell P. Consumer-driven health care: Building partnerships in research. Health Expectations 2005;8(4):352-9.
16. Mills EJ, Montori VM, Ross CP, Shea B, Wilson K, Guyatt GH. Systematically reviewing qualitative studies complements survey design: An exploratory study of barriers to paediatric immunisations. Journal of Clinical Epidemiology 2005;58(11):1101-8.
17. Wells GA, Boers M, Shea B, Brooks PM, Simon LS, Strand CV, Aletaha D, Anderson JJ, Bombardier C, Dougados M, Emery P, Felson DT, Fransen J, Furst DE, Hazes JMW, Johnson KR, Kirwan JR, Landewé RBM, Lassere MND, Michaud K, Suarez-Almazor M, Silman AJ, Smolen JS, Van Der Heijde DMFM, Van Riel PLCM, Wolfe F, Tugwell PS. Minimal disease activity for rheumatoid arthritis: A preliminary definition. Journal of Rheumatology 2005;32(10):2016-24.
18. Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, Hochberg MC, Wells G. Glucosamine therapy for treating osteoarthritis. Cochrane Database of Systematic Reviews (Online) 2005(2).
19. Andersson N, Cockcroft A, Ansari N, Omer K, Losos J, Ledogar RJ, Tugwell P, Shea B. Household cost-benefit equations and sustainable universal childhood immunisation: A randomised cluster controlled trial in south pakistan (ISRCTN12421731). BMC Public Health 2005;5.
20. Tugwell P, Shea B, Boers M, Brooks P, Simon L, Strand V, Wells G. Evidence Based Rheumatology. *BMJ Books* 2004.
21. Brosseau L, Welch V, Wells G, DeBie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (classes I, II and III) for treating osteoarthritis. Cochrane Database of Systematic Reviews (Online) 2004(3).
22. Shea B, Wells G, Cranney A, Cummings S, Sellmeyer D. Review: Calcium supplementation has a small positive effect on bone mineral density but not fractures in postmenopausal women. Evidence-Based Medicine 2004;9(6):170.
23. Shea B, Bonaiuti D, Iovine R, Negrini S, Robinson V, Kemper HC, Wells G, Tugwell P, Cranney A. Cochrane review on exercise for preventing and treating osteoporosis in postmenopausal women. Europa Medicophysica 2004;40(3):199-209.
24. Shea B, Wells G, Cranney A, Zytaruk N, Robinson V, Griffith L, Hamel C, Ortiz Z, Peterson J, Adachi J, Tugwell P, Guyatt G. Calcium supplementation on bone loss in postmenopausal women. Cochrane Database of Systematic Reviews (Online) 2004(1).
25. Tanzer M, Gollish J, Leighton R, Orrell K, Giacchino A, Welsh P, Shea B, Wells G. The effect of adjuvant calcium phosphate coating on a porous-coated femoral stem. Clinical Orthopaedics and Related Research 2004(424):153-60.
26. Brosseau L, Welch V, Wells G, deBie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (classes I, II and III) for treating osteoarthritis. Cochrane Database of Systematic Reviews (Online) 2003(2).
27. Osiri M, Shea B, Robinson V, Suarez-Almazor M, Strand V, Tugwell P, Wells G. Leflunomide for the treatment of rheumatoid arthritis: A systematic review and metaanalysis. Journal of Rheumatology 2003;30(6):1182-90.
28. Wells G, Boers M, Shea B, Anderson J, Felson D, Johnson K, Kirwan J, Lassere M, Robinson V, Simon L, Strand V, Van Riel P, Tugwell P. MCID/low disease activity state workshop: Low disease activity state in rheumatoid arthritis. Journal of Rheumatology 2003;30(5):1110-1.
29. Wells G, Anderson J, Boers M, Felson D, Heiberg T, Hewlett S, Johnson K, Kirwan J, Lassere M, Robinson V, Shea B, Simon L, Strand V, Van Riel P, Tugwell P. MCID/low disease activity state workshop: Summary, recommendations, and research agenda. Journal of Rheumatology 2003;30(5):1115-8.

30. Kirwan J, Heiberg T, Hewlett S, Hughes R, Kvien T, Ahlmèn M, Boers M, Minnock P, Saag K, Shea B, Almazor MS, Taal E. Outcomes from the patient perspective workshop at OMERACT 6. Journal of Rheumatology 2003;30(4):868-72.
31. Tugwell PSL, Qualman A, Judd MG, Dickson E, Frank C, Bombardier C, Davies J, Graham I, Grimshaw J, Hatcher-Roberts J, Hulme J, Jadad A, James P, Jeanes D, Morrice D, Shea B, Sterling L, Wells GA. Knowledge translation of musculoskeletal health research. Journal of Rheumatology 2003;30(3):575-8.
32. Lichtenstein JR, Pope J, Thompson AE, Shea B, Robin V, Fenlon D. Use of sildenafil citrate in raynaud's phenomenon: Comment on the article by thompson et al (11) (multiple letters). Arthritis and Rheumatism 2003;48(1):282-3.
33. Brosseau L, Casimiro L, Milne S, Robinson V, Shea B, Tugwell P, Wells G. Deep transverse friction massage for treating tendinitis. Cochrane Database of Systematic Reviews (Online: Update Software) 2002(4).
34. Robinson V, Tugwell P, Judd M, Shea B, Wells G. Research methodology in rheumatology. Acta Orthopaedica Scandinavica, Supplement 2002;73(305):8-14.
35. Cranney A, Tugwell P, Zytaruk N, Robinson V, Weaver B, Adachi J, Wells G, Shea B, Guyatt G. IV. meta-analysis of raloxifene for the prevention and treatment of postmenopausal osteoporosis. Endocrine Reviews 2002;23(4):524-8.
36. Wells G, Tugwell P, Shea B, Guyatt G, Peterson J, Zytaruk N, Robinson V, Henry D, O'Connell D, Cranney A, Adachi J, Griffith L, McGowan J, Weaver B, Willan A, Rosen CJ, Bilezikian JP, Black DM, Favus MJ, Fitzpatrick LA, Kiel DP, Marcus R, Orwoll ES, Schnitzer TJ. V. meta-analysis of the efficacy of hormone replacement therapy in treating and preventing osteoporosis in postmenopausal women. Endocrine Reviews 2002;23(4):529-39.
37. Cranney A, Wells G, Willan A, Griffith L, Zytaruk N, Robinson V, Black D, Adachi J, Shea B, Tugwell P, Guyatt G. II. meta-analysis of alendronate for the treatment of postmenopausal women. Endocrine Reviews 2002;23(4):508-16.
38. Shea B, Wells G, Cranney A, Zytaruk N, Robinson V, Griffith L, Ortiz Z, Peterson J, Adachi J, Tugwell P, Guyatt G, McGowan J, Weaver B, Willan A, Rosen CJ, Bilezikian JP, Black DM, Favus MJ, Fitzpatrick LA, Kiel DP, Marcus R, Orwoll ES, Schnitzer TJ. VII. meta-analysis of calcium supplementation for the prevention of postmenopausal osteoporosis. Endocrine Reviews 2002;23(4):552-9.
39. Cranney A, Tugwell P, Zytaruk N, Robinson V, Weaver B, Shea B, Wells G, Adachi J, Waldegger L, Guyatt G, Griffith L, McGowan J, Willan A, Rosen CJ, Bilezikian JP, Black DM, Favus MJ, Fitzpatrick LA, Kiel DP, Marcus R, Orwoll ES, Schnitzer TJ. VI. meta-analysis of calcitonin for the treatment of postmenopausal osteoporosis. Endocrine Reviews 2002;23(4):540-51.
40. Cranney A, Tugwell P, Adachi J, Weaver B, Zytaruk N, Papaioannou A, Robinson V, Shea B, Wells G, Guyatt G. III. meta-analysis of risedronate for the treatment of postmenopausal osteoporosis. Endocrine Reviews 2002;23(4):517-23.
41. Guyatt G, Adachi J, Cranney A, Griffith L, McGowan J, Robinson V, Shea B, Tugwell P, Weaver B, Wells G, Willan A, Zytaruk N, Rosen CJ, Bilezikian JP, Black DM, Favus MJ, Fitzpatrick LA, Kiel DP, Marcus R, Orwoll ES, Schnitzer TJ. Meta-analyses of therapies for postmenopausal osteoporosis. Endocrine Reviews 2002;23(4):496-507.
42. Papadimitropoulos E, Wells G, Shea B, Gillespie W, Weaver B, Zytaruk N, Cranney A, Adachi J, Tugwell P, Josse R, Greenwood C, Guyatt G. VIII: Meta-analysis of the efficacy of vitamin D treatment in preventing osteoporosis in postmenopausal women. Endocrine Reviews 2002;23(4):560-9.
43. Bonaiuti D, Shea B, Iovine R, Negrini S, Robinson V, Kemper HC, Wells G, Tugwell P, Cranney A. Exercise for preventing and treating osteoporosis in postmenopausal women. Cochrane Database of Systematic Reviews (Online) 2002(3).

44. Casimiro L, Brosseau L, Robinson V, Milne S, Judd M, Well G, Tugwell P, Shea B. Therapeutic ultrasound for the treatment of rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2002(3).
45. Robinson V, Brosseau L, Casimiro L, Judd M, Shea B, Wells G, Tugwell P. Thermotherapy for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online) 2002(2).
46. Nichol G, McAlister F, Pham B, Laupacis A, Shea B, Green M, Tang A, Wells G. Meta-analysis of randomised controlled trials of the effectiveness of antiarrhythmic agents at promoting sinus rhythm in patients with atrial fibrillation. Heart 2002;87(6):535-43.
47. Brosseau L, Milne S, Robinson V, Marchand S, Shea B, Wells G, Tugwell P. Efficacy of the transcutaneous electrical nerve stimulation for the treatment of chronic low back pain: A meta-analysis. Spine 2002;27(6):596-603.
48. Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. Evaluation and the Health Professions 2002;25(1):116-29.
49. Brosseau L, Casimiro L, Milne S, Robinson V, Shea B, Tugwell P, Wells G. Deep transverse friction massage for treating tendinitis. Cochrane Database of Systematic Reviews (Online: Update Software) 2002(1).
50. Robinson V, Brosseau L, Casimiro L, Judd M, Shea B, Wells G, Tugwell P. Thermotherapy for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online) 2002(1).
51. Tugwell P, Wells G, Peterson J, Welch V, Page J, Davison C, McGowan J, Ramroth D, Shea B. Do silicone breast implants cause rheumatologic disorders? A systematic review for a court-appointed national science panel. Arthritis and Rheumatism 2001;44(11):2477-84.
52. O'Rourke K, Shea B, Wells GA. "Meta-Analysis in Clinical Trials", Applied Statistics in the Pharmaceutical Industry (S. Millard and A. Krause, editors), Springer-Verlag, New York; 2001. p. 397-424.
53. Brosseau L, Casimiro L, Robinson V, Milne S, Shea B, Judd M, Wells G, Tugwell P. Therapeutic ultrasound for treating patellofemoral pain syndrome. Cochrane Database of Systematic Reviews (Online: Update Software) 2001(4).
54. Welch V, Brosseau L, Peterson J, Shea B, Tugwell P, Wells G. Therapeutic ultrasound for osteoarthritis of the knee. Cochrane Database of Systematic Reviews (Online: Update Software) 2001(3).
55. Cranney A, Welch V, Adachi JD, Guyatt G, Krolicki N, Griffith L, Shea B, Tugwell P, Wells G. Etidronate for treating and preventing postmenopausal osteoporosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2001(4).
56. Wells GA, Tugwell P, Brosseau L, Robinson VA, Graham ID, Shea BJ, McGowan J. Philadelphia panel evidence-based clinical practice guidelines on selected rehabilitation interventions: Overview and methodology. Physical Therapy 2001;81(10):1629-40.
57. Kalish RA, McHugh G, Granquist J, Shea B, Ruthazer R, Steere AC. Persistence of immunoglobulin M or immunoglobulin G antibody responses to borrelia burgdorferi 10-20 years after active lyme disease. Clinical Infectious Diseases 2001;33(6):780-5.
58. Thompson AE, Shea B, Welch V, Fenlon D, Pope JE. Calcium-channel blockers for raynaud's phenomenon in systemic sclerosis. Arthritis and Rheumatism 2001;44(8):1841-7.
59. Milne S, Welch V, Brosseau L, Saginur M, Shea B, Tugwell P, Wells G. Transcutaneous electrical nerve stimulation (TENS) for chronic low back pain. Cochrane Database of Systematic Reviews (Online) 2001(2).
60. Welch V, Brosseau L, Shea B, McGowan J, Wells G, Tugwell P. Thermotherapy for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online) 2001(2).
61. Cranney A, Guyatt G, Krolicki N, Welch V, Griffith L, Adachi JD, Shea B, Tugwell P, Wells G. A meta-analysis of etidronate for the treatment of postmenopausal osteoporosis. Osteoporosis International 2001;12(2):140-51.

62. Towheed TE, Anastassiades TP, Shea B, Houpt J, Welch V, Hochberg MC. Glucosamine therapy for treating osteoarthritis. Cochrane Database of Systematic Reviews (Online) 2001(1).
63. Coyle D, Welch V, Shea B, Gabriel S, Drummond M, Tugwell P. Issues of consensus and debate for economic evaluation in rheumatology. Journal of Rheumatology 2001;28(3):642-7.
64. Clinch J, Tugwell P, Wells G, Shea B. Individualized functional priority approach to the assessment of health related quality of life in rheumatology. Journal of Rheumatology 2001;28(2):445-51.
65. Bellamy N, Carr A, Dougados M, Shea B, Wells G. Towards a definition of "difference" in osteoarthritis. Journal of Rheumatology 2001;28(2):427-30.
66. Cranney A, Welch V, Wells G, Adachi J, Shea B, Simon L, Tugwell P. Discrimination of changes in osteoporosis outcomes. Journal of Rheumatology 2001;28(2):413-21.
67. Wells G, Anderson J, Beaton D, Bellamy N, Boers M, Bombardier C, Breedveld F, Carr A, Cranney A, Dougados M, Felson D, Kirwan J, Schiff M, Shea B, Simon L, Smolen J, Strand V, Tugwell P, Van Riel P, Welch VA. Minimal clinically important difference module: Summary, recommendations, and research agenda. Journal of Rheumatology 2001;28(2):452-4.
68. Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, Strand V, Shea B. Looking for important change/differences in studies of responsiveness. Journal of Rheumatology 2001;28(2):400-5.
69. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, Brooks P, Tugwell P. Minimal clinically important differences: Review of methods. Journal of Rheumatology 2001;28(2):406-12.
70. Suarez-Almazor ME, Spooner CH, Belseck E, Shea B. Auranofin versus placebo in rheumatoid arthritis. Schweizerische Rundschau Fur Medizin/Praxis 2001;90(7):264.
71. Haguenauer D, Welch V, Shea B, Tugwell P, Adachi JD, Wells G. Fluoride for the treatment of postmenopausal osteoporotic fractures: A meta-analysis. Osteoporosis International 2000;11(9):727-38.
72. Haguenauer D, Welch V, Shea B, Tugwell P, Wells G. Fluoride for treating postmenopausal osteoporosis. Cochrane Database of Systematic Reviews (Online) 2000(4).
73. Welch V, Brosseau L, Shea B, McGowan J, Wells G, Tugwell P. Thermotherapy for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online) 2000(4).
74. Osiri M, Welch V, Brosseau L, Shea B, McGowan J, Tugwell P, Wells G. Transcutaneous electrical nerve stimulation for knee osteoarthritis. Cochrane Database of Systematic Reviews (Online) 2000(4).
75. Tugwell P, Welch V, Suarez-Almazor M, Shea B, Wells G. Efficacy and toxicity of old and new disease modifying antirheumatic drugs. Annals of the Rheumatic Diseases 2000;59(SUPPL. 1):i32-5.
76. Brosseau L, Welch V, Wells G, Tugwell P, De Bie R, Gam A, Harman K, Shea B, Morin M. Low level laser therapy for osteoarthritis and rheumatoid arthritis: A metaanalysis. Journal of Rheumatology 2000;27(8):1961-9.
77. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Prazosin for raynaud's phenomenon in progressive systemic sclerosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
78. Ortiz Z, Shea B, Suarez Almazor M, Moher D, Wells G, Tugwell P. Folic acid and folinic acid for reducing side effects in patients receiving methotrexate for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
79. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Ketanserin for raynaud's phenomenon in progressive systemic sclerosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
80. Wells G, Haguenauer D, Shea B, Suarez-Almazor ME, Welch VA, Tugwell P. Cyclosporine for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
81. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Cyclophosphamide for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).

82. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Cyclofenil for raynaud's phenomenon in progressive systemic sclerosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
83. Brosseau L, Welch V, Wells G, deBie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (classes I, II and III) for the treatment of osteoarthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
84. Suarez-Almazor ME, Spooner CH, Belseck E, Shea B. Auranofin versus placebo in rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
85. Saenz A, Ausejo M, Shea B, Wells G, Welch V, Tugwell P. Pharmacotherapy for behcet's syndrome. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
86. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Sulfasalazine for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
87. Homik J, Suarez-Almazor ME, Shea B, Cranney A, Wells G, Tugwell P. Calcium and vitamin D for corticosteroid-induced osteoporosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
88. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Methotrexate for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
89. Suarez-Almazor ME, Belseck E, Shea B, Homik J, Wells G, Tugwell P. Antimalarials for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
90. Cranney A, Welch V, Adachi JD, Homik J, Shea B, Suarez-Almazor ME, Tugwell P, Wells G. Calcitonin for the treatment and prevention of corticosteroid-induced osteoporosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
91. Homik J, Cranney A, Shea B, Tugwell P, Wells G, Adachi R, Suarez-Almazor M. Bisphosphonates for steroid induced osteoporosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
92. Brosseau L, Welch V, Wells G, deBie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (classes I, II and III) in the treatment of rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
93. Clark P, Tugwell P, Bennet K, Bombardier C, Shea B, Wells G, Suarez-Almazor ME. Injectable gold for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
94. Criswell LA, Saag KG, Sems KM, Welch V, Shea B, Wells G, Suarez-Almazor ME. Moderate-term, low-dose corticosteroids for rheumatoid arthritis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
95. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Iloprost and cisaprost for raynaud's phenomenon in progressive systemic sclerosis. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
96. Towheed T, Shea B, Wells G, Hochberg M. Analgesia and non-aspirin, non-steroidal anti-inflammatory drugs for osteoarthritis of the hip. Cochrane Database of Systematic Reviews (Online: Update Software) 2000(2).
97. Cranney A, Welch V, Adachi JD, Homik J, Shea B, Suarez-Almazor ME, Tugwell P, Wells G. Calcitonin for the treatment and prevention of corticosteroid-induced osteoporosis. Schweizerische Rundschau Für Medizin/Praxis 2000;89(24):1067-9.
98. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Iloprost and cisaprost for raynaud's phenomenon in progressive systemic sclerosis. Schweizerische Rundschau Für Medizin/Praxis 2000;89(19):828.
99. Pope J, Fenlon D, Thompson A, Shea B, Furst D, Wells G, Silman A. Ketanserin for raynaud's phenomenon in progressive systemic sclerosis. Schweizerische Rundschau Für Medizin/Praxis 2000;89(18):775.

100. Wells G, Haguenauer D, Shea B, Suarez-Almazor ME, Welch VA, Tugwell P. Cyclosporine for rheumatoid arthritis. Schweizerische Rundschau Für Medizin/Praxis 2000;89(17):731.
101. Towheed T, Shea B, Wells G, Hochberg M. Analgesia and non-aspirin, non-steroidal anti-inflammatory drugs for osteoarthritis of the hip. Schweizerische Rundschau Für Medizin/Praxis 2000;89(7):290.
102. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Methotrexate for rheumatoid arthritis. Schweizerische Rundschau Für Medizin/Praxis 1999;88(50):2067.
103. Ortiz Z, Shea B, Suarez-Almazor M, Moher D, Wells G, Tugwell P. Folic acid and folinic acid for reducing side effects in patients receiving methotrexate for rheumatoid arthritis. Schweizerische Rundschau Für Medizin/Praxis 1999;88(50):2068-9.
104. Homik J, Suarez-Almazor ME, Shea B, Cranney A, Wells G, Tugwell P. Calcium and vitamin D for cortico-steroid-induced osteoporosis. Schweizerische Rundschau Für Medizin/Praxis 1999;88(47):1953.
105. Homik JE, Cranney A, Shea B, Tugwell P, Wells G, Adachi JD, Suarez-Almazor ME. A metaanalysis on the use of bisphosphonates in corticosteroid induced osteoporosis. Journal of Rheumatology 1999;26(5):1148-57.
106. Ortiz Z, Shea B, Garcia Dieguez M, Boers M, Tugwell P, Boonen A, Wells G. The responsiveness of generic quality of life instruments in rheumatic diseases. A systematic review of randomized controlled trials. Journal of Rheumatology 1999;26(1):210-6.
107. Nichol G, Dennis DT, Steere AC, Lightfoot R, Wells G, Shea B, Tugwell P. Test-treatment strategies for patients suspected of having lyme disease: A cost-effectiveness analysis. Annals of Internal Medicine 1998;128(1):37-48.
108. Ortiz Z, Shea B, Suarez-Almazor ME, Moher D, Wells GA, Tugwell P. The efficacy of folic acid and folinic acid in reducing methotrexate gastrointestinal toxicity in rheumatoid arthritis. A meta-analysis of randomized controlled trials. Journal of Rheumatology 1998;25(1):36-43.
109. Tugwell P, Dennis DT, Weinstein A, Wells G, Shea B, Nichol G, Hayward R, Lightfoot R, Baker P, Steere AC. Laboratory evaluation in the diagnosis of lyme disease. Annals of Internal Medicine 1997;127(12):1109-23.
110. Wells G, Cranney A, Shea B, Tugwell P. Responsiveness of endpoints in osteoporosis clinical trials. Journal of Rheumatology 1997;24(6):1230-3.
111. Cranney A, Tugwell P, Cummings S, Sambrook P, Adachi J, Silman AJ, Gillespie WJ, Felson DT, Shea B, Wells G. Osteoporosis clinical trials endpoints: Candidate variables and clinimetric properties. Journal of Rheumatology 1997;24(6):1222-9.
112. Cranney A, Tugwell P, Shea B, Wells G. Implications of OMERACT outcomes in arthritis and osteoporosis for Cochrane meta-analysis. Journal of Rheumatology 1997;24(6):1206-7.